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

Anonymous ACL submission

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 factual error. We construct MoFE by combining the experts using weights and logits ensembling strategies and find that the MoFE provides a modular approach to control different factual errors while maintaining performance on standard ROUGE metrics[1].

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.

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 and modular 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.

We use entity overlap and dependency arc entailment (DAE) accuracy (Goyal and 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 and Durrett, 2020, 2021). Additionally, given that experts trained on both entity overlap and DAE metrics try to improve precision and are prone to

[1] Code will be released at https://github.com/anonymous/MoFE

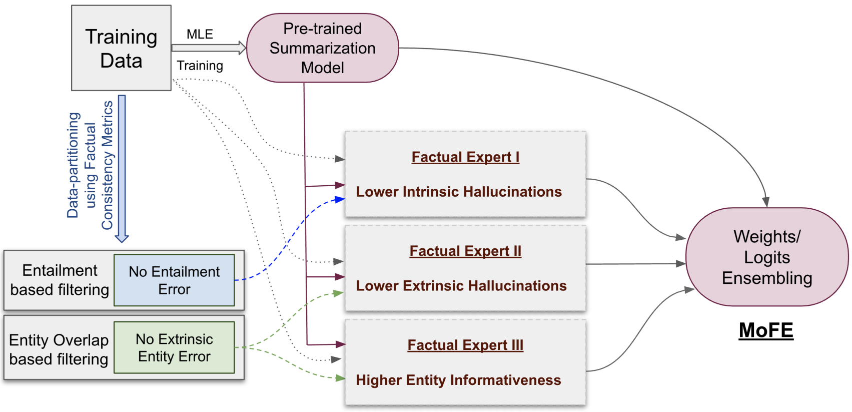

Training Data
MLE
Training
Pre-trained Summarization Model
Data-partitioning using Factual Consistency Metrics
Entailment based filtering
No Entailment Error
Entity Overlap based filtering
No Extrinsic Entity Error
Factual Expert I
Lower Intrinsic Hallucinations
Factual Expert II
Lower Extrinsic Hallucinations
Factual Expert III
Higher Entity Informativeness
Weights/ Logits Ensembling
MoFE

Figure 1: Schematic view of steps for building the MoFE model. First, it uses automated factual consistency metrics to filter out training samples with the desirable factual quality. Then, it trains a factual expert model on the filtered or whole training set and combines them through weights or logits ensembling.

reducing factual recall, MoFE also includes an entity recall-based expert. Subsequently, we combine the above three experts through logits and weights ensembling. We show the schematic view of MoFE in Figure 1.

We evaluate MoFE on two benchmark abstractive summarization datasets in English, 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 models on factual consistency metrics used to train experts, with marginal variations in ROUGE scores. Our empirical results suggest that we can steer the text summarization system to generate faithful content by carefully training expert models. Further, it shows that the benefit of combining multiple experts to control text generation extends beyond broader textual properties such as sentiment and toxicity (as shown by Liu et al. (2021a)), and it can handle constrained text generation with more fine-grained factual qualities.

2 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 precision** 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 and 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 and 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 accuracy**[3] 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 MoFE Model

We propose Mixture of Factual Experts (MoFE) to improve the factual consistency of text summa-

[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.

[3] DAE accuracy measures the number of dependency arcs in summary that are also entailed by the source document.

rization 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 (§3.1). Then, we use reinforcement learning to train expert models for each factual consistency metric (§3.2). Finally, for controlling summary generation, we either directly modify the base model's parameters through weights exsembling (Izmailov et al., 2018) or modify next token probabilities from base model through logits ensembling (§3.3).

3.1 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 and Durrett, 2021) accuracy between the source document and reference summary. We filter all training samples where all of the dependency arcs in the summary are entailed by the source documents.

3.2 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 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, entity precision), 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:

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

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[4] 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[5]. We train experts using the weighted sum of the two losses $\lambda L_\theta^{fc} + (1-\lambda)L_\theta^{kl}$.

$$L_\theta^{kl} = E_x[p_\phi(y^*|x)\ log(p_\phi(y^*|x)\ /\ p_\theta(y^*|x))] \quad (2)$$

Equations 1 and 2 describe the general framework for training our experts. In Eq. 2, we choose y^* depending on the number of factual errors in training samples. 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 (filtered training sam-

[4] 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.

[5] 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 and He, 2021), the benefit of exploration in training text generation systems is limited in the absence of perfect reward functions.

ples from §3.1), we propose to use reference as y^*[6]. On the contrary, 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. Therefore, when factual quality of training data is indeterminable, we propose to use summary sampled following probabilities from then expert (policy) model as y^*.

Intuitively, using reference summary on factually consistent training data is suitable for training experts that aim to improve factual consistency. However, data filtering reduces the number of samples. Given this training data size vs factual quality trade-off, we experiment with both paradigm for training experts. However, with limited compute resources, we recommend performing data filtering followed by RL training to build experts that target content-precision metrics. For recall-related experts, data filtering and mode of RL training is not intuitive and should be empirically determined.

3.3 Mixing Factual Experts

Following the data filtering and RL training steps described in §3.1 and §3.2, we train 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 expert** to maximize recall of salient entities between the generated summary and the reference summary. Note that entity overlap precision is defined with respect to the source document and entity-recall is defined with respect to the reference summary.

Next, we combine the three experts through weights or logits ensembling. We use the element-wise weighted average of all the parameters of pre-trained summarization model and expert models for weights ensembling. For logits ensembling, we use the weighted average of logits from all the experts and the pre-trained model during decoding. The mixing coefficients for all experts and pre-trained models are used to control the factual quality of summaries generated by the ensemble model.

4 Results

4.1 Experimental Setup

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[7] (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 (Zhang et al., 2019b) precision (BS-P) and recall (BS-R) with respect to source and two question answer-based evaluation metrics, FEQA and QuestEval.

4.2 Models

We use the *BART* (Lewis et al., 2020) and *PEGASUS* (Zhang et al., 2019a) released with Huggingface's transformer (Wolf et al., 2020) (*bart-xsum-large*, *pegasus-xsum*, *bart-cnn-large*) as 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 all models during the inference. 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$_W$ and MoFE$_L$ using weights and logits ensembling respectively. Note that experts training targeting specific factual quality may reduce performance on other metrics (e.g. precision-based expert may reduce recall). Therefore, we include an expert in MoFE only if it does not under-perform the base model by more than 5% on any of the DAE-A/S, NER-PS/RT, and ROUGE metrics. An alternative approach could be to stop training when expert's performance falls below a pre-defined threshold.

4.3 Automatic Evaluation

Table 1 summarizes the results on XSUM and CNN/DM datasets. Both MoFE$_W$ and MoFE$_L$ out-

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

[7] 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	BS-P	BS-R	R1	R2	RL
BART XSUM									
Base	76.16	34.75	63.82	**53.66**	88.93	79.86	**45.34**	**22.21**	**37.13**
MoFE$_W$	80.36	**41.08**	66.74	53.20	89.21	79.89	45.00	21.92	36.80
MoFE$_L$	**80.70**	41.06	**66.81**	53.40	**89.24**	**79.94**	45.18	22.03	36.94
PEGASUS XSUM									
Base	73.83	33.22	60.39	**56.39**	88.72	79.68	47.08	**24.54**	**39.29**
MoFE$_W$	75.84	35.36	61.64	56.38	88.81	**79.74**	47.07	24.31	39.11
MoFE$_L$	**75.97**	**35.50**	**61.73**	56.23	**88.82**	**79.74**	**47.12**	24.35	39.16
BART CNN/DM									
Base	96.26	75.0	**98.44**	58.92	93.26	82.62	**44.05**	**21.07**	**40.86**
MoFE$_W$	**96.98**	**77.08**	98.16	60.86	93.30	82.94	44.02	21.02	40.69
MoFE$_L$	96.88	76.01	98.07	**61.79**	**93.35**	**83.12**	43.74	20.86	40.33

Table 1: DAE accuracy, entity precision, entity recall and ROUGE scores for the base and MoFE models on XSUM and CNN/DM datasets.

perform BART and PEGASUS across all factual consistency metrics on the XSUM dataset. Similarly, both models outperform BART on CNN/DM dataset with marginal degradation on entity precision (NER-P). 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). Next, neither of the MoFE models lowers ROUGE scores substantially on either of the XSUM or CNN/DM datasets, the worst being 0.53 drop on ROUGE-L for MoFE$_L$ on CNN/DM. We also find that MoFE models improve BERTScore precision (BS-P) and recall (BS-R) with respect to the source article on both XSUM and CNN/DM datasets. This is particularly interesting given recent work on benchmarking different evaluation metrics suggests that BERTScore precision with respect to the source document correlates with the human judgment of factuality (Pagnoni et al., 2021).

Between logits and weights ensembling, we find both performing comparably 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. Accordingly, for fair comparison with the base model, we use MoFE$_W$ for all our analyses.

Model	XSUM		CNN/DM	
	FEQA	QEval	FEQA	QEval
Base	25.77	36.54	**38.22**	59.24
MoFE$_W$	**27.87**	37.32	35.85	59.79
MoFE$_L$	27.74	**37.43**	34.64	**59.90**

Table 2: QA metrics-based evaluations of BART and corresponding MoFE models.

QA-based Evaluations: In table 2, we report results for BART and corresponding MoFE models on QA-based metrics. MoFE models improve on the QA-based QuestEval metric on both XSUM and CNN/DM datasets. However, both MoFE$_W$ and MoFE$_L$ 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?”).*

4.4 Human Evaluation

Following Cao and Wang (2021), we perform pairwise comparison of summaries, where a human

Model	DAE-A	DAE-S	NER-PS	NER-RT	R1	R2	RL
BART XSUM							
Base	76.16	34.75	63.82	53.66	**45.34**	22.21	**37.13**
Unfiltered-MLE	75.22	33.48	62.63	**54.23**	45.27	**22.28**	37.09
Filtered-MLE	78.86	39.04	66.14	52.20	44.96	21.93	36.91
MoFE$_W$	**80.36**	**41.08**	**66.74**	53.20	45.00	21.92	36.80
BART CNN/DM							
Base	96.26	75.0	**98.44**	58.92	44.05	21.07	40.86
Unfiltered-MLE	95.19	67.44	97.72	**61.93**	**44.28**	**21.23**	**40.88**
Filtered-MLE	96.96	**77.20**	98.21	60.91	44.05	21.11	40.71
MoFE$_W$	**96.98**	77.08	98.16	60.86	44.02	21.02	40.69

Table 3: Ablation: DAE accuracy, entity precision, entity recall and ROUGE scores for different weight-ensembled models on XSUM and CNN/DM datasets.

annotator rate each MoFE$_W$ generated summary against the BART summary for factual consistency. We use randomly sampled 100 articles from each of the XSUM and CNN/DM datasets. First, two annotators independently annotated 20 articles-summaries pairs from XSUM to calculate inter-annotator agreement[8]. We found that two annotations achieve high Krippendoeff's alpha coefficient (Krippendorff, 2011) of 0.83847. Then, one annotator rated remaining 80 XSUM and 100 CNN/DM articles. Annotators found MoFE$_W$ improves (degrades) factual consistency on 30% (11%) summaries on XSUM data, and improves (degrades) factual consistency on 5% (1%) summaries on CNN/DM data. Factual consistency remained unchanged for remaining 59% and 94% summaries from XSUM and CNN/DM datasets respectively. Given higher percentage of factual errors as well as higher empirical gain on XSUM data, we further analyze 30 XSUM summaries from MoFE and BART models using SummVis (Vig et al., 2021) tool. We discuss our findings in appendix, §B.

5 Analysis

5.1 Effects of Data Filtering and RL Training

In Table 3, we evaluate how training data filtering and RL-based training contribute to the improved performance of MoFE. *Unfiltered-MLE* is an ensemble of four BART models, including the best performing *base*, and *Filtered-MLE* is an ensemble of experts, trained exclusively with the MLE loss on corresponding filtered data, and the base model. First, we find that ensembling multiple BART models improves ROUGE scores and NER recall, but not factual consistency metrics defined by DAE accuracy and NER precision. On the other hand, *Filtered-MLE* ensemble consistently outperforms both *Base* and *Unfiltered-MLE* models on factual consistency metrics, underlining the importance of using factually correct samples during training. MoFE$_W$ model, that is based on RL training to directly optimize factual consistency, further improves the performance on XSUM data. However, on CNN/DM data, MoFE$_W$ and *Filtered-MLE* perform comparably. To further understand the reasons for different behavior, we analyze summaries sampled using the probabilities from BART models trained on XSUM and CNN/DM datasets. As shown in Table 12 and 13 in appendix, we find that XSUM-BART model-sampled summaries exhibit varied factual behavior, generating both factually consistent and inconsitent summaries. On the other hand, CNN-BART model sampled summaries are overly extractive and mainly differ on the sentences sampled from source article but not on factual consistency. Evidently, benefit from RL training can be pre-inferred by analyzing summaries sampled from the baseline models used to initialize the policy.

	All		Filtered	
	DAE-A	DAE-S	DAE-A	DAE-S
BART	76.67	35.79	76.67	35.79
Reference	75.55	31.33	**82.53**	**44.09**
Model	**84.1**	**46.92**	80.27	41.70

Table 4: Performance of DAE experts trained with reference and sampled summary (Model)-based KL loss on all training data and filtered subset of training data.

KL divergence loss vs training data quality: In Table 4, we report the validation performance of DAE expert trained using reference summary and model-sampled summary on filtered XSUM training subset and whole XSUM training data. We observe that both variants of experts improve performance on DAE-A/S metrics when trained on the

[8] We also crowd-sourced annotations using Amazon Mechanical Turk, but we found inter-annotator agreement to be extremely low (Krippendoeff's alpha coefficient < 0.10). Based on crowd-sourced annotations, MoFE$_W$ improves factual consistency on 22% (31%), and degrades factual consistency on 10% (21%) of CNN/DM (XSUM) summaries.

filtered subset. However, the margin of improvement is higher for reference-based experts, implying the advantage of minimizing KL divergence on reference 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 metrics. On contrary, we find experts minimizing KL divergence on sampled summary effective, 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 *reiterate that factual quality of training data affects the performance of experts*. On factually consistent samples, we can use either of the reference or sampled summary to define KL divergence loss. However, when samples contain factual errors, reference summary may not be effective.

5.2 Mixture of experts vs joint RL training

Model	DAE-A	DAE-S	NER-PS	NER-RT
BART	76.16	34.75	63.82	53.66
DAE	**83.83**	**46.83**	**69.09**	51.82
NER-P	76.81	36.02	67.37	53.69
NER-R	75.48	33.56	63.50	**55.04**
Joint	80.74	41.33	68.71	51.78
MoFE_W	80.36	41.08	66.74	53.20

Table 5: DAE accuracy, entity precision and entity recall of individual experts on XSUM data.

In Table 5, we compare performance of individual experts and an RL model trained to jointly optimize all rewards. First, all three experts outperform the BART model, on their respective factual consistency metric. Importantly, DAE expert performs better than (or comparable to) NER-P expert on NER-PS metric. Dependency arc error subsumes extrinsic entity error as dependency arcs corresponding to extrinsic entities can not be entailed by the source document. We consider this a desirable behavior given we do not need to train multiple experts if we can choose the right set of reward function/ metric.

The *Joint* model that uses average of DAE, NER-P and NER-R rewards and trains on data filtered according to all three metrics, perform slightly better than MoFE_W on DAE-A/S and NER-PS metric. However, it obtains 1.42 points lower entity recall as well as performs consistently worse than the DAE expert across all metrics. Notably, MoFE_W has the flexibility to include multiple experts and adjust for degradation in performance on any metric by including an appropriate expert during the decoding time, as discussed in the next section. Therefore, joint model can also be used as a new expert in MoFE and resulting degradation in NER recall can be adjusted by the NER-R expert.

5.3 Effects of Mixing Coefficients on Ensemble of an Expert and BART

We combine each expert and the BART model with different mixing coefficients (α) and plot their performance on XSUM validation data in Figure 2. We use weights ensembling for our analyses and evaluate models on DAE-A/S and NER-PS/RT metrics. First, we find that the performance of the ensemble of expert and BART model on the respective metric roughly lies on the linear line connecting the performance of the individual expert and BART models. On the metrics that are not part of expert training, we find that the performance of the ensemble model either remains approximately unchanged (e.g. DAE-A, NER-PS metrics for the NER-R expert) or lies on the linear line (e.g. NER-PS/RT metrics for the DAE expert). Given the linear dependence, we can decide the mixing coefficient for an expert depending on the tolerance value for the ensemble model on all metrics. Further, we can compensate for the reduction in performance of the ensemble model on any metric by training an expert targeting that specific metric. For instance, to compensate for the reduction in performance of the ensemble of DAE and BART on the NER-RT metric, we can add an NER-R expert that obtains higher NER recall than the base BART model. Note that, the modular characteristics of MoFE also allows us to choose different values of mixing coefficients for each of the experts and BART model depending on the significance of different factual errors in the target application.

6 Related Work

Factual consistency metrics and analysis Abstractive text summarization metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019b) 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 such as entailment, question answering-based evaluation (Falke

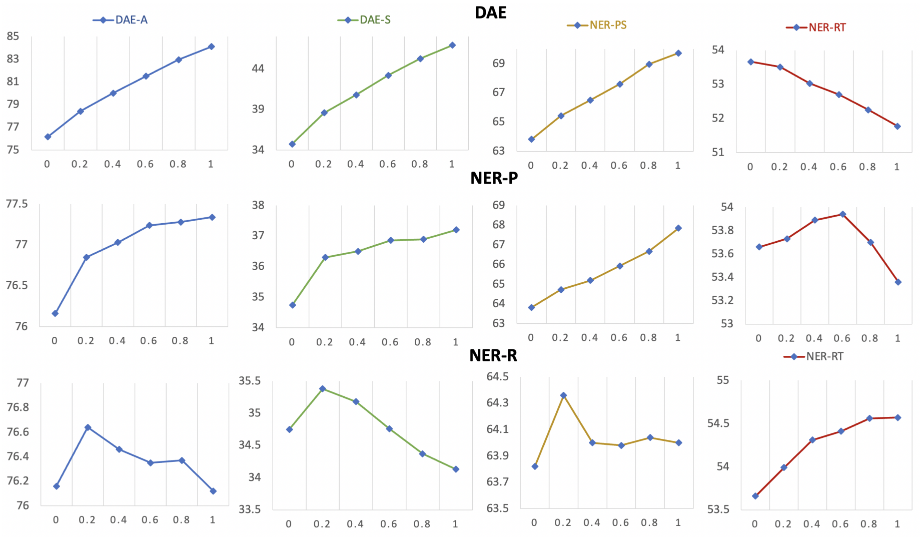


Figure 2: Variations in the performance of weight-ensembled expert and BART models with different values of mixing coefficient α (α=0.0 corresponds to only BART model, and α = 1.0 corresponds to only expert model.).

et al., 2019; Kryscinski et al., 2020; Maynez et al., 2020; Zhou et al., 2021; Eyal et al., 2019; Scialom et al., 2019; Wang et al., 2020; Durmus et al., 2020; Scialom et al., 2021), etc. The slew of work on factual evaluation metrics has also given rise to research focused on comparing different metrics (Gabriel et al., 2021; Fabbri et al., 2021; Pagnoni et al., 2021; Goyal and 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 while Pagnoni et al. (2021) concluded that entailment-based FactCC exhibits the highest correlations with the human judgment of factuality. 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.

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), loss truncation (Kang and Hashimoto, 2020), 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.

7 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 MLE-trained 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. Further, our results and analyses highlight that text generation can be controlled for fine-grained factual qualities at decoding time through appropriately trained experts.

References

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

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)*, pages 6251–6258, Online. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. 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.

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

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

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, Online. Association for Computational Linguistics.

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

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. 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*, pages 1449–1462, Online. Association for Computational Linguistics.

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

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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

Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics.

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

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 9332–9346, Online. Association for Computational Linguistics.

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

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

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021a. 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)*, pages 6691–6706, Online. Association for Computational Linguistics.

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

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

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

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

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. 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*, pages 2727–2733, Online. Association for Computational Linguistics.

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

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

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

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

S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195, Los Alamitos, CA, USA. IEEE Computer Society.

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

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Priyam Tejaswin, Dhruv Naik, and Pengfei Liu. 2021. How well do you know your summarization datasets? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3436–3449, Online. Association for Computational Linguistics.

Jesse Vig, Wojciech Kryscinski, Karan Goel, and Nazneen Rajani. 2021. 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*, pages 150–158, Online. Association for Computational Linguistics.

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

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

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. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

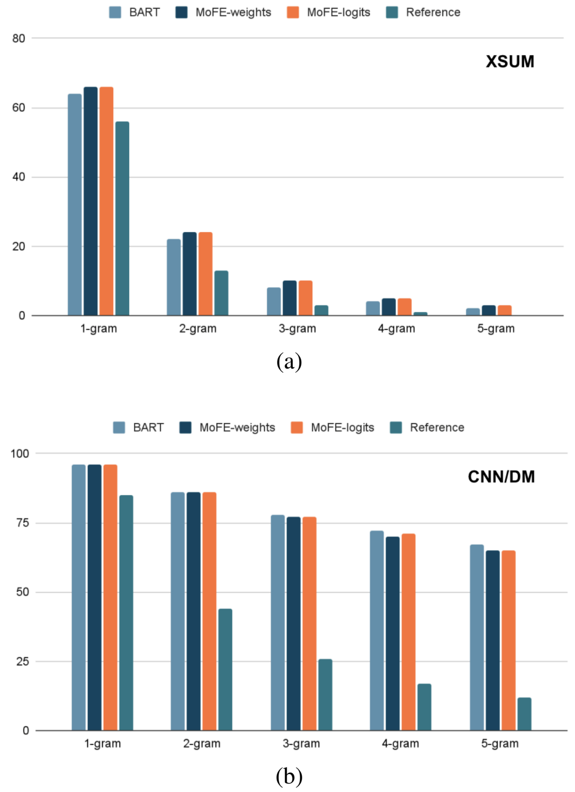
Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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*, pages 1393–1404, Online. Association for Computational Linguistics.

A Extractiveness vs Faithfulness

	XSUM		CNN/DM	
	AL	UO	AL	UO
Reference	21.10	11.87	55.01	46.25
BART	18.84	12.25	64.32	61.80
MoFE$_W$	18.54	12.30	70.04	67.34
MoFE$_L$	18.92	12.62	73.29	70.62

Table 6: Average length (AL) and average number of common unigrams (AO) between summary and source document for reference summaries and BART and MoFE models.



(a)

(b)

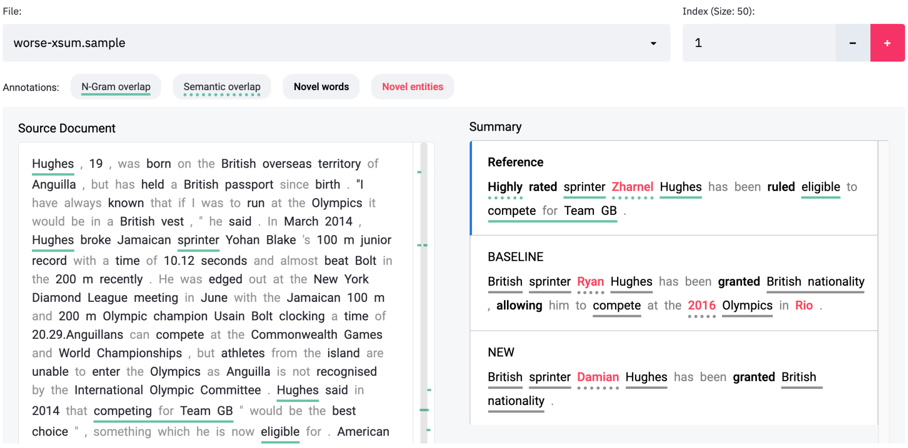
Figure 3: Percentage of overlapped n-grams in XSUM and CNN/DM summaries.

We analyze the extractiveness-faithfulness trade-off for the BART and MoFE models. We compare the ratio of n-grams in summaries that appear in the source document in Figure 3, and average length of summaries and average number of common unigrams between summaries and source documents in Table 6. All BART and MoFE models are highly extractive on CNN/DM datasets and they tend to generate summaries longer than the references. Also, the difference in n-grams overlap percentage between reference and model-based summaries is much higher on the CNN/DM data. On the contrary, models generate shorter summaries than the references on XSUM data, but they still generate summaries with higher n-grams overlap percentages. It is generally observed that neural models, including BART, tend to increase the extractiveness (Durmus et al., 2020).

Both MoFE$_W$ and MoFE$_L$ increase the average number of overlapped unigrams on both XSUM and CNN/DM datasets. Further, MoFE models increase the average summary length on CNN/DM. This is expected given 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). This is likely to exacerbate the difference between the length of model-generated and reference summaries. It is worth noting that logits ensembling increases the length of generated summary more than the weights ensembling on both XSUM and CNN/DM datasets, another disadvantage of the former besides an increase in decoding time. Overall, 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.

B SummVis Analysis

We analyze 30 samples from each of the MoFE$_L$ and BART models on XSUM data using *SummVis* (Vig et al., 2021) tool. We show 8 interesting samples from the analyzed 30. 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. 4 and 5; *II)* removes all factual errors, Fig. 6; *III)* replaces one factual error with another, Fig. 7; *IV)* adds factual error, Fig. 8; and *V)* adds or removes world knowledge, Fig. 9 and 10. 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.



worse-xsum.sample

1

Source Document

Hughes , 19 , was born on the British overseas territory of Anguilla , but has held a British passport since birth . "I have always known that if I was to run at the Olympics it would be in a British vest , " he said . In March 2014 , Hughes broke Jamaican sprinter Yohan Blake 's 100 m junior record with a time of 10.12 seconds and almost beat Bolt in the 200 m recently . He was edged out at the New York Diamond League meeting in June with the Jamaican 100 m and 200 m Olympic champion Usain Bolt clocking a time of 20.29.Anguillans can compete at the Commonwealth Games and World Championships , but athletes from the island are unable to enter the Olympics as Anguilla is not recognised by the International Olympic Committee . Hughes said in 2014 that competing for Team GB " would be the best choice " , something which he is now eligible for . American

Summary

Reference

Highly rated sprinter Zharnel Hughes has been ruled eligible to compete for Team GB .

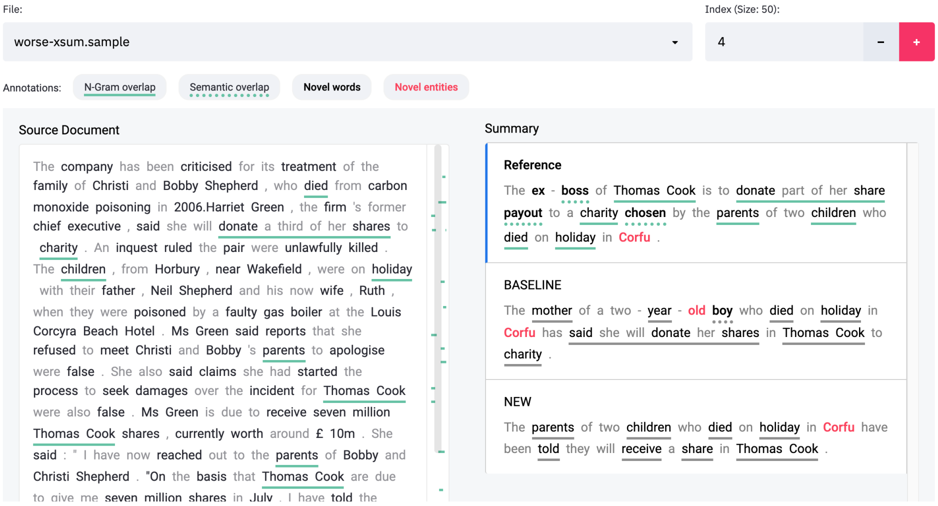
BASELINE

British sprinter Ryan Hughes has been granted British nationality , allowing him to compete at the 2016 Olympics in Rio .

NEW

British sprinter Damian Hughes has been granted British nationality .

(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"*.



worse-xsum.sample

4

Source Document

The company has been criticised for its treatment of the family of Christi and Bobby Shepherd , who died from carbon monoxide poisoning in 2006.Harriet Green , the firm 's former chief executive , said she will donate a third of her shares to charity . An inquest ruled the pair were unlawfully killed . The children , from Horbury , near Wakefield , were on holiday with their father , Neil Shepherd and his now wife , Ruth , when they were poisoned by a faulty gas boiler at the Louis Corcyra Beach Hotel . Ms Green said reports that she refused to meet Christi and Bobby 's parents to apologise were false . She also said claims she had started the process to seek damages over the incident for Thomas Cook were also false . Ms Green is due to receive seven million Thomas Cook shares , currently worth around £ 10m . She said : " I have now reached out to the parents of Bobby and Christi Shepherd . "On the basis that Thomas Cook are due to give me seven million shares in July . I have told the

Summary

Reference

The ex - boss of Thomas Cook is to donate part of her share payout to a charity chosen by the parents of two children who died on holiday in Corfu .

BASELINE

The mother of a two - year - old boy who died on holiday in Corfu has said she will donate her shares in Thomas Cook to charity .

NEW

The parents of two children who died on holiday in Corfu have been told they will receive a share in Thomas Cook .

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

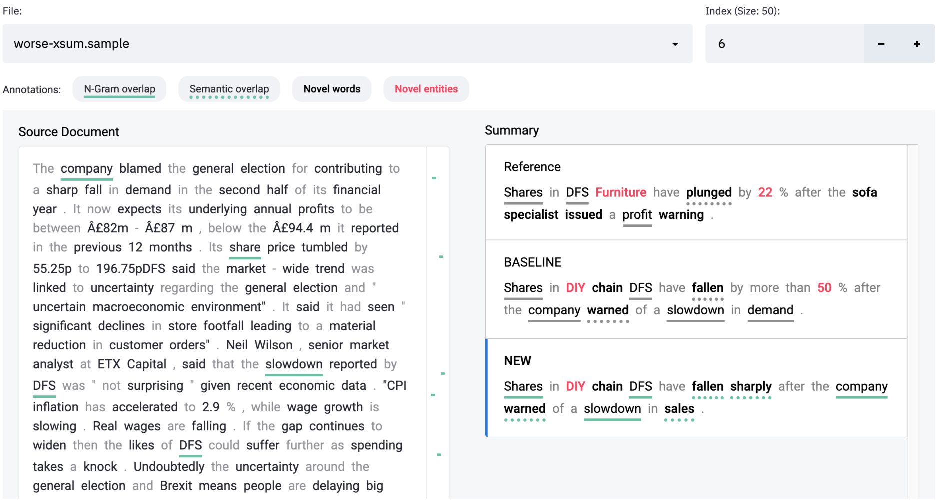
C Experimental Details

C.1 Models

We use the *BART* (Lewis et al., 2020) and *PEGASUS* (Zhang et al., 2019a) released with Huggingface's transformer (Wolf et al., 2020) (*bart-xsum-large*, *pegasus-xsum*, *bart-cnn-large*) as 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 all models during the inference, e.g. beam size of 6 (4), minimum and maximum sequence length of 11 (56) and 62 (142), etc. for the XSUM (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 models 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, β_1: 0.9, β_2: 0.999, 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



File: worse-xsum.sample

Index (Size: 50): 6

Source Document

The company blamed the general election for contributing to a sharp fall in demand in the second half of its financial year . It now expects its underlying annual profits to be between Â£82m - Â£87 m , below the Â£94.4 m it reported in the previous 12 months . Its share price tumbled by 55.25p to 196.75pDFS said the market - wide trend was linked to uncertainty regarding the general election and " uncertain macroeconomic environment" . It said it had seen " significant declines in store footfall leading to a material reduction in customer orders" . Neil Wilson , senior market analyst at ETX Capital , said that the slowdown reported by DFS was " not surprising " given recent economic data . "CPI inflation has accelerated to 2.9 % , while wage growth is slowing . Real wages are falling . If the gap continues to widen then the likes of DFS could suffer further as spending takes a knock . Undoubtedly the uncertainty around the general election and Brexit means people are delaying big

Summary

Reference

Shares in DFS Furniture have plunged by 22 % after the sofa specialist issued a profit warning .

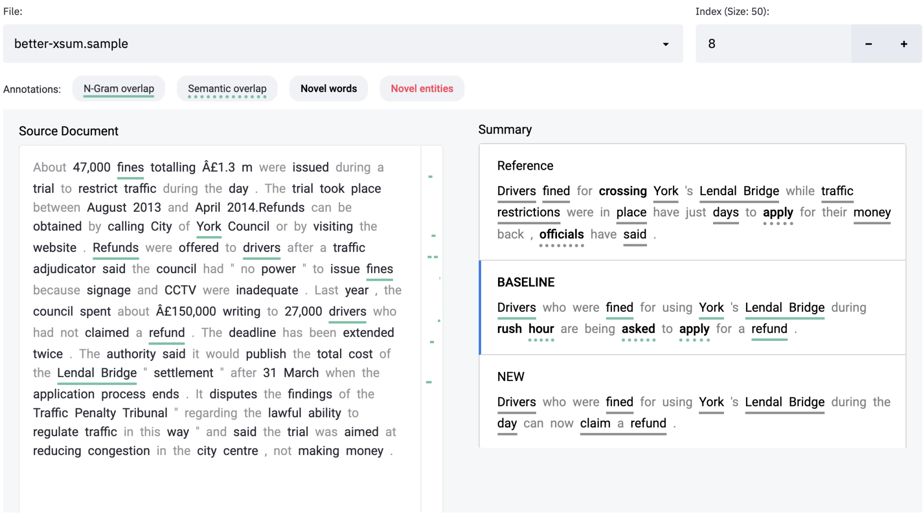
BASELINE

Shares in DIY chain DFS have fallen by more than 50 % after the company warned of a slowdown in demand .

NEW

Shares in DIY chain DFS have fallen sharply after the company warned of a slowdown in sales .

Figure 5: In this example, BART hallucinates percentage amount (50%). MoFE replaces percentage amount to a generic word *sharply*. Both BART and MoFE hallucinates DIY.



File: better-xsum.sample

Index (Size: 50): 8

Source Document

About 47,000 fines totalling Â£1.3 m were issued during a trial to restrict traffic during the day . The trial took place between August 2013 and April 2014.Refunds can be obtained by calling City of York Council or by visiting the website . Refunds were offered to drivers after a traffic adjudicator said the council had " no power " to issue fines because signage and CCTV were inadequate . Last year , the council spent about Â£150,000 writing to 27,000 drivers who had not claimed a refund . The deadline has been extended twice . The authority said it would publish the total cost of the Lendal Bridge " settlement " after 31 March when the application process ends . It disputes the findings of the Traffic Penalty Tribunal " regarding the lawful ability to regulate traffic in this way " and said the trial was aimed at reducing congestion in the city centre , not making money .

Summary

Reference

Drivers fined for crossing York 's Lendal Bridge while traffic restrictions were in place have just days to apply for their money back , officials have said .

BASELINE

Drivers who were fined for using York 's Lendal Bridge during rush hour are being asked to apply for a refund .

NEW

Drivers who were fined for using York 's Lendal Bridge during the day can now claim a refund .

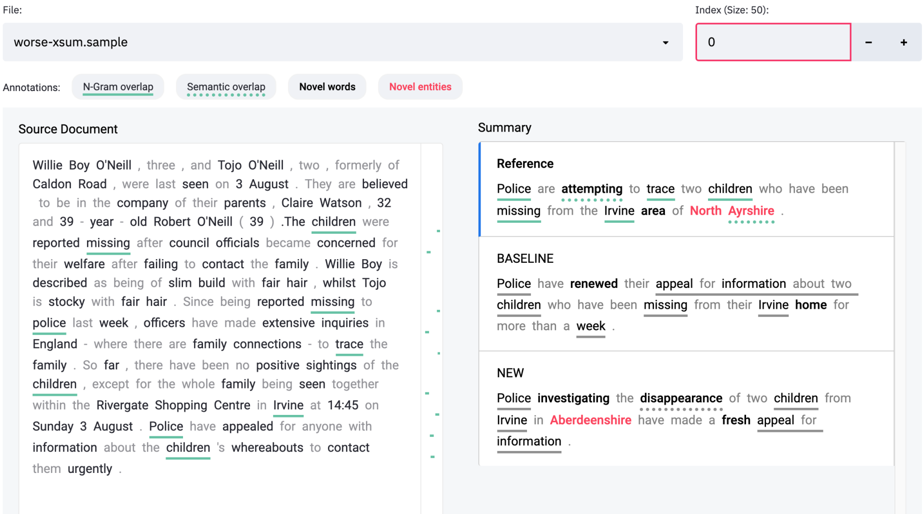
Figure 6: In this example, BART hallucinates *rush hour*. In contrast, MoFE generates factually correct summary.

reference (NER-R). We construct two variants of MoFE, MoFE$_W$ and MoFE$_L$ using weights and logits ensembling respectively. Note that 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 mixing coefficients (α^i) for weight ensembling 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 and NER-R experts from the MoFE given they degraded DAE-S accuracy by greater than 5%. Similar to XSUM, we use grid search to find mixing coefficients for CNN data, but we assigned a minimum weight of 0.2 to the DAE expert and the BART model. In our analyses, however, we found that mixing coefficients can be intuitively guessed based on the performance of individual experts and the desired performance of MoFE on different evaluation metrics. In Table 7, we report the constituents experts for each of the MoFE models and datasets.

	DAE	NER-P	NER-R
XSUM BART	Model sampled, All data	Model sampled, All data	Model sampled, All data
XSUM PEGASUS	Reference, Filtered data	Reference, Filtered data	Model sampled, All data
CNN/DM BART	Reference, Filtered data	NA	NA

Table 7: Constituents Models in MoFE.

In Tables 8, 9, 10 and 11, we report all our re-



worse-xsum.sample

0

Source Document

Willie Boy O'Neill , three , and Tojo O'Neill , two , formerly of Caldon Road , were last seen on 3 August . They are believed to be in the company of their parents , Claire Watson , 32 and 39 - year - old Robert O'Neill (39) .The children were reported missing after council officials became concerned for their welfare after failing to contact the family . Willie Boy is described as being of slim build with fair hair , whilst Tojo is stocky with fair hair . Since being reported missing to police last week , officers have made extensive inquiries in England - where there are family connections - to trace the family . So far , there have been no positive sightings of the children , except for the whole family being seen together within the Rivergate Shopping Centre in Irvine at 14:45 on Sunday 3 August . Police have appealed for anyone with information about the children 's whereabouts to contact them urgently .

Summary

Reference

Police are attempting to trace two children who have been missing from the Irvine area of North Ayrshire .

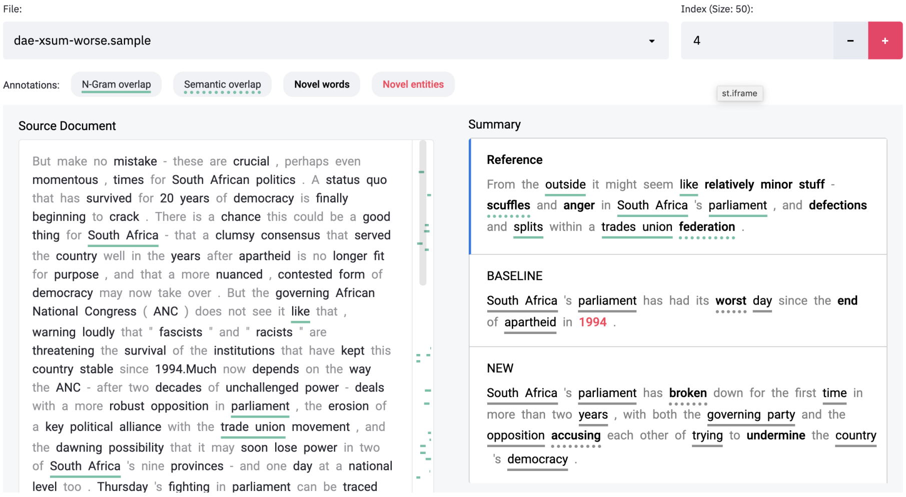
BASELINE

Police have renewed their appeal for information about two children who have been missing from their Irvine home for more than a week .

NEW

Police investigating the disappearance of two children from Irvine in Aberdeenshire have made a fresh appeal for information .

Figure 7: Both BART and MoFE generate different factual errors, BART hallucinates *more than a week* and MoFE hallucinates *Aberdeenshire*.



dae-xsum-worse.sample

4

Source Document

But make no mistake - these are crucial , perhaps even momentous , times for South African politics . A status quo that has survived for 20 years of democracy is finally beginning to crack . There is a chance this could be a good thing for South Africa - that a clumsy consensus that served the country well in the years after apartheid is no longer fit for purpose , and that a more nuanced , contested form of democracy may now take over . But the governing African National Congress (ANC) does not see it like that , warning loudly that " fascists " and " racists " are threatening the survival of the institutions that have kept this country stable since 1994.Much now depends on the way the ANC - after two decades of unchallenged power - deals with a more robust opposition in parliament , the erosion of a key political alliance with the trade union movement , and the dawning possibility that it may soon lose power in two of South Africa 's nine provinces - and one day at a national level too . Thursday 's fighting in parliament can be traced

Summary

Reference

From the outside it might seem like relatively minor stuff - scuffles and anger in South Africa 's parliament , and defections and splits within a trades union federation .

BASELINE

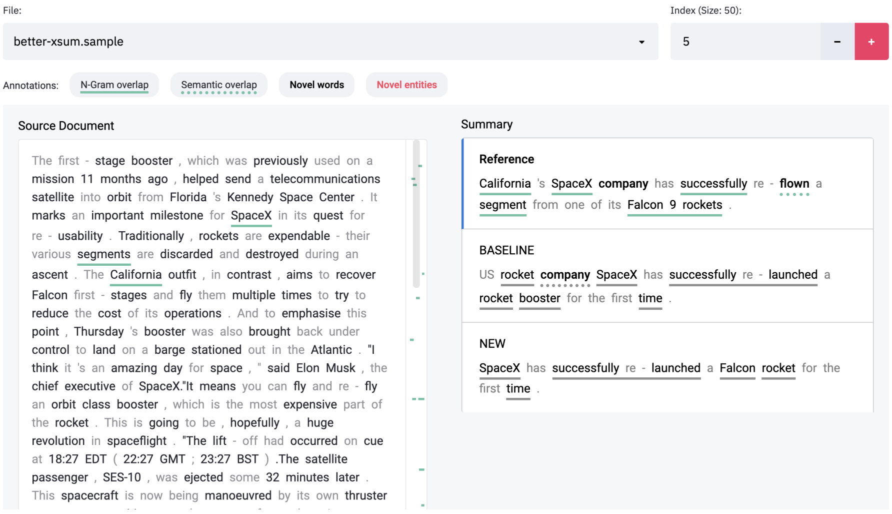
South Africa 's parliament has had its worst day since the end of apartheid in 1994 .

NEW

South Africa 's parliament has broken down for the first time in more than two years , with both the governing party and the opposition accusing each other of trying to undermine the country 's democracy .

Figure 8: In this example, BART incorporates world knowledge *"end of apartheid"* and is factually consistent otherwise. MoFE adds factual error *"two years"*.

sults.



File:

better-xsum.sample

Index (Size: 50):

5

Annotations: N-Gram overlap Semantic overlap Novel words Novel entities

Source Document

The first - stage booster , which was previously used on a mission 11 months ago , helped send a telecommunications satellite into orbit from Florida 's Kennedy Space Center . It marks an important milestone for SpaceX in its quest for re - usability . Traditionally , rockets are expendable - their various segments are discarded and destroyed during an ascent . The California outfit , in contrast , aims to recover Falcon first - stages and fly them multiple times to try to reduce the cost of its operations . And to emphasise this point , Thursday 's booster was also brought back under control to land on a barge stationed out in the Atlantic . "I think it 's an amazing day for space , " said Elon Musk , the chief executive of SpaceX."It means you can fly and re - fly an orbit class booster , which is the most expensive part of the rocket . This is going to be , hopefully , a huge revolution in spaceflight . "The lift - off had occurred on cue at 18:27 EDT (22:27 GMT ; 23:27 BST) .The satellite passenger , SES-10 , was ejected some 32 minutes later . This spacecraft is now being manoeuvred by its own thruster

Summary

Reference

California 's SpaceX company has successfully re - flown a segment from one of its Falcon 9 rockets .

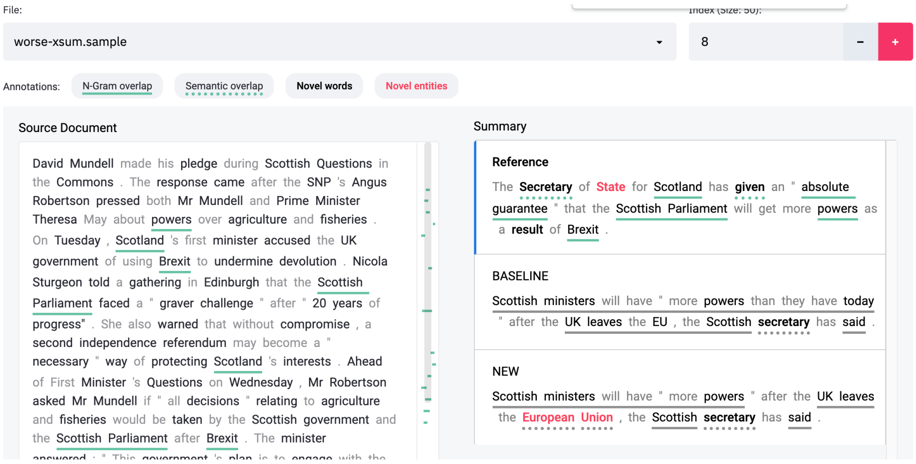
BASELINE

US rocket company SpaceX has successfully re - launched a rocket booster for the first time .

NEW

SpaceX has successfully re - launched a Falcon rocket for the first time .

Figure 9: 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).



File:

worse-xsum.sample

Index (Size: 50):

8

Annotations: N-Gram overlap Semantic overlap Novel words Novel entities

Source Document

David Mundell made his pledge during Scottish Questions in the Commons . The response came after the SNP 's Angus Robertson pressed both Mr Mundell and Prime Minister Theresa May about powers over agriculture and fisheries . On Tuesday , Scotland 's first minister accused the UK government of using Brexit to undermine devolution . Nicola Sturgeon told a gathering in Edinburgh that the Scottish Parliament faced a " graver challenge " after " 20 years of progress" . She also warned that without compromise , a second independence referendum may become a " necessary " way of protecting Scotland 's interests . Ahead of First Minister 's Questions on Wednesday , Mr Robertson asked Mr Mundell if " all decisions " relating to agriculture and fisheries would be taken by the Scottish government and the Scottish Parliament after Brexit . The minister answered : " This government 's plan is to engage with the

Summary

Reference

The Secretary of State for Scotland has given an " absolute guarantee " that the Scottish Parliament will get more powers as a result of Brexit .

BASELINE

Scottish ministers will have " more powers than they have today " after the UK leaves the EU , the Scottish secretary has said .

NEW

Scottish ministers will have " more powers " after the UK leaves the European Union , the Scottish secretary has said .

Figure 10: Both BART and MoFE are factually correct, though MoFE replaces *EU* with *European Union* (world knowledge).

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
Unfiltered-MLE$_W$	75.22	33.48	62.63	54.23	45.27	22.28	37.09
Unfiltered-MLE$_L$	75.02	35.53	63.01	52.57	45.14	22.41	37.38
RL Models							
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$_W$	80.36	41.08	66.74	53.20	45.00	21.92	36.80
MoFE$_L$	80.70	41.06	66.81	53.40	45.18	22.03	36.94
MLE Trained Models							
DAE-MLE	80.52	38.83	68.43	51.96	44.84	21.41	36.38
NER-MLE	78.79	36.11	66.42	53.32	44.78	21.33	36.24
Filtered-MLE$_W$	78.86	39.04	66.14	52.20	44.96	21.93	36.91
Filtered-MLE$_L$	78.65	41.61	66.45	51.07	44.86	22.13	37.17

Table 8: DAE accuracy, entity precision, entity recall and ROUGE scores of BART-based models 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
Unfiltered-MLE$_W$	95.11	66.99	97.46	63.16	43.86	20.98	40.37
Unfiltered-MLE$_L$	95.11	67.0	97.46	63.16	43.86	20.98	40.38
RL Models							
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	95.11	67.45	98.23	61.06	44.43	21.36	41.25
MoFE$_W$	96.98	77.08	98.16	60.86	44.02	21.02	40.69
MoFE$_L$	96.88	76.01	98.07	61.79	43.74	20.86	40.33
MLE Trained Models							
DAE-MLE	97.12	78.18	98.27	60.13	44.11	21.15	40.90
NER-MLE	95.36	67.29	98.23	60.74	44.44	21.31	41.23
Filtered-MLE$_W$	96.96	77.20	98.21	60.91	44.05	21.11	40.71
Filtered-MLE$_L$	96.94	77.11	98.07	61.71	43.73	20.88	40.31

Table 9: DAE accuracy, entity precision, entity recall and ROUGE scores of BART-based models on CNN/DM test set.

Model	DAE-A	DAE-S	NER-PS	NER-RT	R1	R2	RL
BART	73.83	33.22	60.39	56.39	47.08	24.54	**39.29**
DAE	**76.71**	**36.47**	**62.03**	56.10	47.03	24.19	39.01
NER-P	75.20	34.03	61.83	55.41	46.76	24.01	38.84
NER-R	73.21	33.62	60.59	**56.84**	46.76	**24.55**	39.27
MoFE$_W$	75.84	35.36	61.64	56.38	47.07	24.31	39.11
MoFE$_L$	75.97	35.50	61.73	56.23	**47.12**	24.35	39.16

Table 10: DAE accuracy, entity precision, entity recall and ROUGE scores of PEGASUS-based models on XSUM test set.

	DAE				NER-P		NER-R	
	All		Filtered		All	Filtered	All	Filtered
	DAE-A	DAE-S	DAE-A	DAE-S	NER-PS	NER-PS	NER-RT(-PS)	NER-RT(-PS)
BART	76.67	35.79	76.67	35.79	64.30	64.30	53.55 (64.30)	53.55 (64.30)
Reference	75.55	31.33	**82.53**	**44.09**	60.87	**69.06**	44.47(60.88)	51.27(68.33)
Model	**84.1**	**46.92**	80.27	41.70	**67.84**	66.88	**54.57(63.95)**	**53.79(65.60)**

Table 11: 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. NER-R expert trained with reference-based KL divergence loss perform worse than the one trained with sampled summary-based KL divergence loss on NER-RT metric. But note that NER-R is not a metric to measure the factual consistency and performance of NER-R expert on NER-PS (a factual consistency metric) is as expected.

Source	"Prosecutors say managers at Peanut Corporation of America shipped peanuts and products they knew were tainted. More than 575 people in more than 40 states were sickened in the outbreak, including hundreds of children. A lawyer for former owner Stewart Parnell said inspectors had been aware of the company's testing practices. The charges carry maximum penalties of 20 years in prison, prosecutors say. Mr Parnell, his brother, former Vice-President Michael Parnell, as well as former plant managers Samuel Lightsey and Daniel Kilgore, have been charged with fraud, selling ""adulterated and misbranded food"", and conspiracy. In addition, former plant worker Mary Wilkerson was charged with obstruction of justice. Kilgore has already pleaded guilty in the case. Prosecutors say the Parnells, Mr Lightsey and Kilgore conspired to manufacture and sell peanuts and peanut products that lab tests had shown were tainted with salmonella. They created fake certificates saying the foods were safe, when in fact they had either not been tested or had been found to have been contaminated, prosecutors said. Stewart Parnell, Mr Lightsey and Ms Wilkerson lied to visiting government inspectors, they said. "When those responsible for producing or supplying our food lie and cut corners, as alleged in the indictment, they put all of us at risk," said Stuart Delery, head of the justice department's civil division. "The Department of Justice will not hesitate to pursue any person whose criminal conduct risks the safety of Americans who have done nothing more than eat a peanut butter and jelly sandwich.
Sample 1 (Inconsistent)	**Four** former executives of a US peanut company have been charged in connection with an outbreak of salmonella that **killed more than 500 people**.
Sample 2 (Inconsistent)	**Three** former executives of a US peanut company have been charged in connection with an outbreak of salmonella that **killed more than 500 people**.
Sample 3 (Consistent)	Former executives of a US peanut company have been charged in connection with an outbreak of salmonella that left hundreds of people sick.

Table 12: Summaries sampled following probabilities from the XSUM-BART model. The model is abstractive, and sampled summaries differ in factual qualities.

Source	(CNN)The only thing crazier than a guy in snowbound Massachusetts boxing up the powdery white stuff and offering it for sale online? People are actually buying it. **For $89, self-styled entrepreneur Kyle Waring will ship you 6 pounds of Boston-area snow in an insulated Styrofoam box** – enough for 10 to 15 snowballs, he says. But not if you live in New England or surrounding states. **"We will not ship snow to any states in the northeast!" says Waring's website, ShipSnowYo.com.** "We're in the business of expunging snow!" **His website and social media accounts claim to have filled more than 133 orders for snow – more than 30 on Tuesday alone**, his busiest day yet. **With more than 45 total inches, Boston has set a record this winter for the snowiest month in its history.** Most residents see the huge piles of snow choking their yards and sidewalks as a nuisance, but Waring saw an opportunity. According to Boston.com, it all started a few weeks ago, when Waring and his wife were shoveling deep snow from their yard in Manchester-by-the-Sea, a coastal suburb north of Boston. He joked about shipping the stuff to friends and family in warmer states, and an idea was born. His business slogan: "Our nightmare is your dream!" At first, ShipSnowYo sold snow packed into empty 16.9-ounce water bottles for $19.99, but the snow usually melted before it reached its destination. So this week, Waring began shipping larger amounts in the Styrofoam cubes, which he promises will arrive anywhere in the U.S. in less than 20 hours. He also has begun selling a 10-pound box of snow for $119. Many of his customers appear to be companies in warm-weather states who are buying the snow as a gag, he said. Whether Waring can sustain his gimmicky venture into the spring remains to be seen. But he has no shortage of product. "At this rate, it's going to be July until the snow melts," he told Boston.com. "But I've thought about taking this idea and running with it for other seasonal items. Maybe I'll ship some fall foliage."
Sample 1	Kyle Waring sells 6 pounds of Boston-area snow in an insulated Styrofoam box for $89. His website and social media accounts claim to have filled more than 133 orders for snow. With more than 45 total inches, Boston has set a record this winter for the snowiest month in its history.
Sample 2	Self-styled entrepreneur Kyle Waring will ship 6 pounds of Boston-area snow in an insulated Styrofoam box for $89. "We will not ship snow to any states in the northeast!" says Waring's website, ShipSnowYo.com. With more than 45 total inches, Boston has set a record this winter for the snowiest month in its history.
Sample 3	Self-styled entrepreneur Kyle Waring is shipping 6 pounds of Boston-area snow in an insulated Styrofoam box for $89. His website and social media accounts claim to have filled more than 133 orders for snow – more than 30 on Tuesday alone. With more than 45 total inches, Boston has set a record this winter for the snowiest month in its history.

Table 13: Summaries sampled following probabilities from the CNN/DM-BART model. The model is extremely extractive, and sampled summaries differ mainly on the sentences selected from source article.