based on the deletions, additions and kept n-grams[5] with respect to the original sentence.[6] For human evaluation of the model's outputs, 20% of the evaluation dataset was used. Crowd-workers were provided with the model outputs and the corresponding supposably consistent claims. They were instructed to score the model outputs from 1 to 5 (1 being the poorest and 5 the highest), on grammaticality and agreement with the claim.

Table 1 reports the automatic and human evaluation results. Our model gets the highest SARI score, showing that it is the closest to humans in modifying the text for the corresponding tasks. Humans also score our outputs the highest for consistency with the claim, an essential criterion of our task. In addition, the outputs are more grammaticality sound compared to those from other methods.

Examining the gold answers, we notice that many of them include very minimal and local modifications, keeping much of the original sentence. The M. Concat model keeps most of the original sentence as is, even at the cost of being inconsistent with the claim. This corresponds to a high KEEP score but a lower SARI score overall, and a low human score on supporting the claim. Claim Ext. and Paraphrase do not maintain the structure of the original sentence, and perform poorly on KEEP, leading to a low SARI score. The Split-no-Copy model has the same low ADD score as Claim Ext. since instead of copying the accurate information from the claim, it generates other tokens.

Data Augmentation For 41850 $\boldsymbol{Dis}$ pairs in the FEVER training data, our method generates synthetic evidence sentences leading to 41850 $\boldsymbol{Agr}$ pairs. We train the BERT fact-checking classifier with this augmented data and report the performance on the symmetric dataset in Table 2. In addition, we repeat the human evaluation process on the generated augmentation pairs and report it in Table 1.

Our method's outputs are effective for augmentation, outperforming a classifier trained only on the original biased training data by an absolute 1.7% on the TEST set and an absolute 3.0% on the +TURK set. The outputs of the Paraphrase and Copy Claim baselines are not Wikipedia-like, making them ineffective for augmentation. All the baseline approaches augment the false claims with a supported evidence. However, the success of our method in producing supporting evidence while trying to maintain a Wikipedia-like structure, leads to more effective augmentations.

Masker Analysis To evaluate the performance of the masker model, we test its capacity to modify $\boldsymbol{Agr}$ and $\boldsymbol{Dis}$ pairs from the FEVER development set to a neutral relation. We measure the accuracy of the pretrained classifier in predicting neutral versus the percentage of masked words from the sentence. For a finer evaluation, we manually annotated 75 $\boldsymbol{Agr}$ and 76 $\boldsymbol{Dis}$ pairs with the minimal required mask for neutrality and compute the per token F1 score of the masker against them.

λ	ACC	SIZE	Δ	PREC	REC	F1
.5	5.1	0.0	5	0.0	0.0	0.0
.4	80.0	26.3	**54**	27.2	75.1	**39.9**
.3	77.0	27.5	50	25.9	71.6	38.0
.2	81.6	31.1	51	23.1	74.8	35.3

Table 3: Results of different values of λ for the masker with syntactic regularization. The left three columns describe the accuracy and average mask size (% of the sentence) over the FEVER development set with the masked evidence and a neutral target label. Δ is ACC $-$ SIZE. The right three columns contain the precision, recall and F1 of the masks that we have human annotations for. For results without syntactic regularization see the appendix.

The results for different values of the regularization coefficient are reported in Table 3. Increasing the regularization coefficient helps to minimize the mask size and to improve the precision while maintaining the classifier accuracy and the mask recall. However, setting λ too large, can collapse the solution to no masking at all. The generation experiments use the outputs of the $\lambda = 0.4$ model.

6 Conclusion

In this paper, we introduce the task of automatic fact-guided sentence modification. Given a claim and an old sentence, we learn to rewrite it to produce the updated sentence. Our method overcomes the challenges of this conditional generation task by breaking it into two steps. First, we identify the polarizing components in the original sentence and mask them. Then, using the residual sentence and the claim, we generate a new sentence which is consistent with the claim. Applied to a Wikipedia fact update evaluation set, our method successfully generates correct Wikipedia sentences using the guiding claims. Our method can also be used for data augmentation, to alleviate the bias in fact verification datasets without any external data, reducing the relative error by 13%.

7 Acknowledgments

We thank the anonymous reviewers and the MIT NLP group for their helpful discussion and comments. This work is supported by DSO grant DSOCL18002.

Illum vel libero animi id ut error dolorem culpa, ex nihil placeat sequi dolore alias mollitia molestias eaque ipsum dolorum, pariatur cum velit asperiores voluptas quam corrupti sint molestiae facere, labore facilis atque facere repellendus explicabo, amet asperiores unde nemo repellat id nam dolorum incidunt magnam nulla.Deserunt quo consequuntur repudiandae voluptas tenetur dolorem voluptate doloremque explicabo debitis ex, minus nobis nemo eos, nisi possimus quasi corporis laborum?Reprehenderit earum dolores mollitia, explicabo iste odit ducimus quod perspiciatis eum ipsam voluptas optio, quaerat eius impedit ex cum sunt laboriosam voluptatem ducimus qui ullam, sequi vel libero cor-

[5]We use the default up to 4-grams setting.

[6]Following (**?**) we use the F1 measure for all three sets, including deletions. The final SARI score is the geometric mean of the ADD, DEL and KEEP score.

rupti?Architecto fugit impedit laboriosam, iure earum natus voluptate vitae fugit praesentium laboriosam tempora, veniam amet incidunt sint