Generating personalized article edits on collaborative editing platforms

Anonymous ACL submission

Abstract

NLP methods to generate edits on collaborative editing platforms can help users to edit more efficiently and suggest locations within an article for editing. Existing methods have largely ignored the *personalized* aspect of editing–the diverse styles, interests, and editing intentions that affect user edits. In this paper, we analyze two personalization methods: augmenting models with user behavior clusters and user tags. We demonstrate that these methods, when combined with a new architecture, generate edits that are closer to ground-truth Wikipedia edits when compared to an existing strong baseline. Our experiments test edits for both edit type (insertion or deletion) and word choice, and include a user study collecting feedback from human evaluators. Finally, we introduce a new dataset of Wikipedia edits to facilitate future innovation.

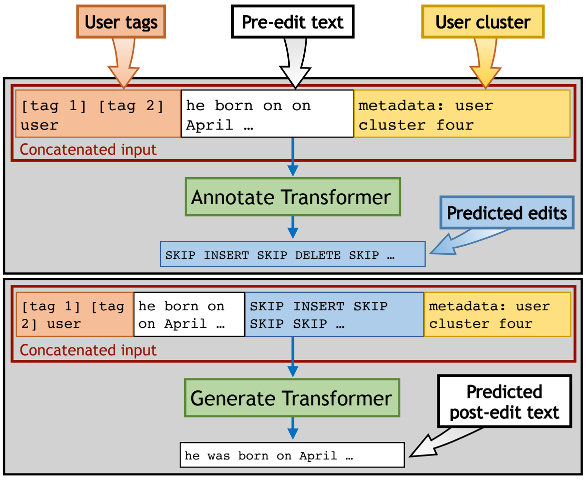

Figure 1: Annotate-Generate (AG) model consisting of two sub-models: an annotator to predict edits and a generator to generate the post-edit text. Also includes user tag and user cluster personalization.

1 Introduction

Neural NLP methods for generating edits on collaborative editing platforms such as Wikipedia are useful for a range of practical tasks, such as assisting users to make article edits efficiently through predictive text, suggesting locations in an article where a user might want to make an edit, and auditing existing article edits for anomalies. However, existing work overlooks the importance of generating *personalized* edits.

To see why personalization matters, consider two users: one tends to clean up articles by removing bad citations, while the other user focuses on adding new up-to-date information to articles. The editor model should tend to predict more removed words for the first user, and more inserted words for the second user. Additionally, each user has a unique writing style and a tendency to focus on particular topics, and an editor model should be able to capture this.

We consider two types of personalization. One method augments models with features obtained from user clustering based on previous user actions (number of additions, deletions, etc.), while the other adds user tags which allow the neural network to learn individual user styles during fine-tuning.

Additionally, we consider two general types of personalized model: End-to-End (E2E) models, which directly predict the post-edit text given the pre-edit text, and Annotate-Generate (AG) models, which split the task into two phases: the prediction of edits (insertions, deletions, etc.) and the generation of post-edit text given these edit predictions. Personalization is relevant for both phases.

Our experiments show that the Full model, which is an AG model supplemented with both user behavioral clusters and user tags, outperforms the baseline unpersonalized E2E model.

2 Related Work

Some existing work, such as Botha et al. (2018) and Miltner et al. (2019) has focused on making restricted edits, rather than general edits. Other work, such as Lebret et al. (2016), Iso et al. (2020)

and Faltings et al. (2020), generates edits given additional side information. Yin et al. (2019) represents edits in high-dimensional space, allowing for clustering and searching of edits, but not generating an edit given only the pre-edit text. Of the above methods, only Miltner et al. (2019), a refactoring tool that suggests repetitive edits based on past behavior for a given user, includes any form of personalization.

Apart from research in edit generation, some work instead classifies or models edits to study Wikipedia. Examples include Yang et al. (2017) and Marrese-Taylor et al. (2019).

The basic problem of predicting edits shares some similarity with non-parametric language models (Guu et al., 2018; Khandelwal et al., 2020; He et al., 2020), which generate text by first selecting and then making edits to a candidate text.

None of these papers considers the generation of personalized and general text edits to Wikipedia articles, which is the intent of this paper.

3 Models

We investigate two general types of model: End-to-End (E2E) models and Annotate-Generate (AG) models. Both model types can be augmented with two forms of personalization: user behavior clusters and user tags.

3.1 Model personalization

User tag personalization works by randomly choosing two unique words from the vocabulary for every user. These are passed into our models as text prepended to the model's other text input, before the separator word "user". This is inspired by Mireshghallah et al. (2021), who showed that user prefix embeddings[1] were outperformed by user tags for sentiment analysis. Our initial experiments with prefixes were also outperformed by user tags.

In addition to these user tags, we also consider a method of clustering users based on their observed behavior in the training set. We create histograms for each user, whose bars correspond to the percentage of words skipped, inserted, replaced, or deleted across all training examples for that user. We then cluster users based on these histograms using the birch algorithm.[2] We obtained good results with a threshold of 0.01 and 16 clusters. We postpend user cluster information to the input text in the following form: "metadata: user cluster [#]", where [#] is the cluster number as a word, such as "fourteen".

See Figure 1 for a visual representation of user cluster and user tag personalization. Performance did not change significantly if the relative order of input text, user tags, user cluster information, and predicted label sequences was altered, as long as this order remained consistent.

3.2 End-to-end (E2E) models

Given an edit $\mathbf{x}_-^{(i)} \rightarrow \mathbf{x}_+^{(i)}$, where $\mathbf{x}_-^{(i)}$ is the pre-edit text and $\mathbf{x}_+^{(i)}$ is the post-edit text, an end-to-end model directly models

$$p(\mathbf{x}_+^{(i)}|\mathbf{x}_-^{(i)}, t(u), c(u); \theta), \qquad (1)$$

where θ represents the neural network parameters, $t(u)$ indicates optional user tag personalization, and $c(u)$ indicates optional user cluster personalization (see Section 3.1). Because we have an input sequence and an output sequence, this task calls for a sequence to sequence model. Sequence to sequence models are typically used for machine translation, but in this case the input and output languages are the same. Our initial experiments considered an LSTM (Zhong et al., 2019), but we found that a deep Transformer architecture performed better (Vaswani et al., 2017). In particular, we use a T5 Transformer model (Raffel et al., 2020), with twelve hidden layers, all of which are fine-tuned during training (see Section 3.4). Each hidden layer consists of 768 dimensional hidden states and 12-head attention mechanisms.[3]

3.3 Annotate-Generate (AG) models

In contrast to the E2E models, the AG model consists of two sub-models. The first model is the annotator model, which models

$$p(\Delta(\mathbf{x}^{(i)})|\mathbf{x}_-^{(i)}, t(u), c(u); \theta), \qquad (2)$$

where $\Delta(\mathbf{x})$ is an edit label sequence–a sequence of the words "SKIP", "INSERT", "DELETE", and "REPLACE", which indicates a shortest-distance word edit between the pre-edit text $\mathbf{x}_-^{(i)}$ and post-edit text $\mathbf{x}_+^{(i)}$ in terms of Levenshtein distance (modified to compute the distance in terms of word edits, rather than character edits). The second model, the

[1] These user prefix embeddings are similar to prefix-tuning (Li and Liang, 2021) embeddings, but with full fine-tuning.

[2] Sklearn implementation (Pedregosa et al., 2011).

[3] For the T5 task prefix, we use the phrase "edit encyclopedia article: ".

generator, takes as input the predicted edit label sequence from the annotator model ($\hat{\Delta}(\mathbf{x}^{(i)})$) and models the post-edit text distribution:

$$p(\mathbf{x}_+^{(i)}|\mathbf{x}_-^{(i)}, \hat{\Delta}(\mathbf{x}^{(i)}), t(u), c(u); \theta). \qquad (3)$$

Both the annotator and generator models are based on T5 models, with the same general architecture as E2E models (see Section 3.2).[4] The generator model receives the predicted edit label sequence as postpended text added to the input string $\mathbf{x}_-^{(i)}$ after the separator word "metadata" (see Figure 1).

One motivation for splitting the edit generation task into two phases is to to decouple the two tasks of predicting user edits and generating the post-edit text. Another motivation is that conditioning generation on specific edit actions can help discourage models from simply predicting that the post-edit text and pre-edit text are the same (see Appendix A for examples of this).

3.4 Training and data preparation

All of our models are fine-tuned versions of the HuggingFace (Wolf et al., 2020) t5-base model.[5] We train the annotate and generate models in two separate stages. For all models, we use cross entropy loss and Adam opitimization. After each training epoch, we evaluate on the validation set. After 15 epochs of training, we choose the model with the lowest validation loss.

We fine-tune models using the dataset described in Section 4, but we first filter out examples that resulted in no change after tokenization, or that result in the complete deletion of the pre-edit text. We also filtered out all users with fewer than 45 editing examples, due to an inherent limitation of user tag features that requires a significant number of past edits. We then split the dataset into 80% training, 10% evaluation, and 10% test subsets. We also ensure that edits for each user are roughly distributed 80%/10%/10% across these three subsets.

AG annotator sub-models are fine-tuned using the training set ground truth edit labels (Section 3.1), with a batch size of six. E2E models and AG generator sub-models models are fine-tuned using the training set ground-truth post-edit text, with a batch size of four.

Because of the way our dataset is constructed (see Section 4), most words between the pre- and post-edit text remain the same, which results in a large number of SKIPs in the edit label sequence $\Delta(\mathbf{x}^{(i)})$. To encourage the model to focus on learning substantive changes, we experimented with removing all SKIPs from the edit label sequence. This resulted in similar or slightly improved performance, so this is the version we use in experiments.

4 Dataset

We also introduce a new dataset, AmericanPoliticians, which consists of edit data from English Wikipedia articles within the category "American Politicians". For each article, we considered up to 500 of the most recent historical edits, but filtered out edits made by users with fewer than 50 edits. For each individual edit, we found the locations of all changes made within the page using diff software and treated each separate change location as a unique data example. We restrict the length of pre- and post-edit strings to be at most 100 words. If edits are longer than 100 words, we discard them; if shorter than 100 words, we include available surrounding context to bring the text length up to a maximum of 100 words. This results in a total of 298,582 individual edits from 33,769 articles, edited by 7,439 individual users.

Potential limitations of this dataset include the fact that it is only in English, and that the way we construct examples means that edits will be more likely to come near the center of the example.

5 Results

Table 1 shows the performance for different E2E and AG models on the test set. It considers the following models:

- **E2E (Baseline Model)**: An E2E model without personalization.
- **E2E-c**: An E2E model augmented with user behavioral cluster personalization.
- **E2E-t**: An E2E model augmented with randomized user tags.
- **E2E-ct**: An E2E model augmented with both user clusters and user tags.
- **AG**: An AG model without personalization.
- **AG-c**: An AG model augmented with user behavioral cluster personalization.
- **AG-t**: An AG model augmented with randomized user tags.

[4] For the annotator T5 task prefix, we use the phrase "predict encyclopedia edits: ", while for the generator prefix we use "edit encyclopedia article: ".

[5] HuggingFace has an Apache 2.0 License and is intended for NLP derivative works such as this one.

Table 1: Measures of performance across different models on the test set. See Section 5 for an explanation of these measures and models.

Model	**Bleu$_1$**	**Recall$_+$**	**Recall$_-$**
E2E	0.310	0.365	0.901
E2E-c	0.311	0.356	0.900
E2E-t	0.315	0.353	0.875
E2E-ct	0.323	0.349	0.898
AG	0.309	0.337	0.900
AG-c	0.318	0.356	**0.912**
AG-t	0.306	0.336	0.904
AG-ct	**0.334**	**0.377**	**0.912**

- **AG-ct (Full Model)**: An AG model augmented with both user clusters and user tags.

Because the AG-ct model includes both types of personalization and the AG architecture, which are the novel contributions of this paper, we call this model the Full Model. In contrast, the E2E model lacks these, so it serves as our Baseline Model. Our user study (Section 5.1) and Appendix A compare this Full Model against the Baseline model. The metrics displayed in Table 1 are as follows:

- **Bleu$_1$**: a unigram bleu score comparing the predicted post-edit text vs. the actual post-edit text. Crucially, this measure considers only the set of words that were actually changed from the pre-edit text–that is, words that are inserted, deleted, or replaced. In the case of a replaced word, we include both the replaced word and its replacement in this set. Performing a Bleu score over all words is inappropriate for our dataset, since the majority of words per edit remain unchanged for each example.
- **Recall$_+$** (**Recall$_-$**): the recall rate of insertions (deletions) where an edit is considered an insertion (deletion) if the post-edit word count count increased (decreased).

We see in Table 1 that the Full Model (AG-ct), which combines the AG architecture with both types of personalization, performs the best on all of metrics. However, to verify that the AG-ct actually result in higher-quality edits as judged by humans, we perform a user study to further compare the Full Model against the Baseline model.

5.1 User study

For our user study, we randomly chose 500 unique examples from the test set for which the Baseline model (E2E) and the Full Model (AG-ct) produced non-identical edits. We labeled the ground truth post-edit text as the "reference text" and the outputs of the two models as the two "candidate texts", and asked users to "note the difference between each candidate text and the reference text" and to "choose the closer candidate". If the user "absolutely can't decide" between the candidates, they could vote that the two candidates were "about the same". We performed the user study using Amazon's Mechanical Turk. We required users to have a Master's Qualification from Amazon.

Table 2: Results of our user study."% majority" indicates examples for which each method obtained a majority consensus (at least 2/3 votes); while "% unanimous" indicates 3/3 votes. "About the same" is a consensus that the two methods performed "about the same".

	% majority	% unanimous
Baseline (E2E)	34%	16%
Full (AG-ct)	**41%**	**20%**
About the same (Ties)	16% (9%)	5% (N/A)

For each example, we required three votes by three different users. The results of the user study appear in Table 2, and indicate that users preferred edits generated by the Full Model over those generated by the Baseline model.

To verify that the performance of the Full Model over the Baseline is statistically significant, we performed a bootstrap significance test (Berg-Kirkpatrick et al., 2012) of 100,000 bootstrap samples and obtained a p-value of 0.0492. For each bootstrap sample, we first drew 500 examples randomly with replacement, and then for each example drew three votes randomly with replacement.

6 Conclusion

In this work, we motivate the need to include personalization in neural editor models when generating edits on collaborative editing platforms, and we introduce two personalization methods, along with an AG model for this task.

We show that our Full Method outperforms a non-personalized baseline, based on metrics which test generated edits for edit type (insertion or deletion) as well as word choice (bleu score). In addition, human evaluators have chosen our method over a non-personalized Baseline. Finally, we introduce a new dataset to facilitate future work.

7 Ethical considerations

As with many text generation methods, we recognize that there are potential risks with our edit generation model. Such risks include the production of malicious edits that are undetectable; or a user's editing style could be copied to produce edits that impersonate that user in a malicious way. These dangers are somewhat mitigated by our use of Wikipedia data, since Wikipedia has mechanisms in place to prevent vandalism of its edits (protecting articles, blocking malicious users, etc). We do not feel our method would allow malicious users to more easily transgress these defenses.

Another potential concern with work such as this is the privacy of users. However, the only individuals we expect to be mentioned by name in our dataset are public figures such as the American politicians from whose Wikipedia articles we collected data. Although we include the usernames of Wikipedia editors in our dataset, these usernames are voluntarily created, mostly anonymous, and publicly available on Wikipedia, so we do not feel that these users' privacy is any further infringed by our work.

When performing our user study, we did not inform users how the data obtained from their votes would be used, which we recognize potentially introduces a means for this data to be used in a way to which the user would not offer consent. However, users did have a means to contact us, ask questions, and raise concerns. None of these voluntary participants expressed any such concerns about how this data would be used.

We also recognize the environmental impact of training deep neural models. For this reason, we provide here an estimate of the total required computational budget for developing our method. Our models were trained using two NVIDIA RTX2080 GPUs, and we estimate a total of 30 days worth of computation on these two GPUs, which includes a search for hyperparameters and initial training of models with alternative personalization methods such as prefix embeddings (see Section 3.1). The expense of computation also means that our main experiments of Table 1 are single-run experiments, rather than means of multiple-run experiments with variance information.

References

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005.

Jan A Botha, Manaal Faruqui, John Alex, Jason Baldridge, and Dipanjan Das. 2018. Learning to split and rephrase from wikipedia edit history. *arXiv preprint arXiv:1808.09468*.

Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. 2020. Text editing by command.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *Transactions of the Association for Computational Linguistics*, 6:437–450.

Junxian He, Taylor Berg-Kirkpatrick, and Graham Neubig. 2020. Learning sparse prototypes for text generation.

Hayate Iso, Chao Qiao, and Hang Li. 2020. Fact-based text editing. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. *arXiv preprint arXiv:1603.07771*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Edison Marrese-Taylor, Pablo Loyola, and Yutaka Matsuo. 2019. An edit-centric approach for wikipedia article quality assessment.

Anders Miltner, Sumit Gulwani, Vu Le, Alan Leung, Arjun Radhakrishna, Gustavo Soares, Ashish Tiwari, and Abhishek Udupa. 2019. On the fly synthesis of edit suggestions. *Proc. ACM Program. Lang.*, 3(OOPSLA).

Fatemehsadat Mireshghallah, Vaishnavi Shrivastava, Milad Shokouhi, Taylor Berg-Kirkpatrick, Robert Sim, and Dimitrios Dimitriadis. 2021. Useridentifier: Implicit user representations for simple and effective personalized sentiment analysis.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

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. Huggingface's transformers: State-of-the-art natural language processing.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010.

Pengcheng Yin, Graham Neubig, Miltiadis Allamanis, Marc Brockschmidt, and Alexander L. Gaunt. 2019. Learning to represent edits.

Guoqiang Zhong, Xin Lin, Kang Chen, Qingyang Li, and Kaizhu Huang. 2019. Long short-term attention.

A Comparison of generated edits

Below we include ten examples from the test set. Each example includes the ground truth edit made by the user, the edit made by the Full Model (AG-ct), and the edit made by the Baseline Model (E2E). We also include brief comments for each of these edits.

Model	Edit	Comment
Ground truth	restoration renovation and professional management of the historic loews jersey theatre . [citation] [citation] ~~these tax abatements have not been without controversy . jersey city residents have criticized these pilot agreements for depriving the public schools of funds and disproportionately saddling owners of non piloted property with local taxes . [citation] republicans from outside of jersey city have criticized fulop for these pilot agreements since piloted buildings do not pay school taxes and thus sustain an artificially high level of state aid . [citation]~~ in jersey city was ranked by atlantic magazine as the th most	Large deletion in middle of example
Full	(Identical to ground truth edit)	Correct edit
Baseline	restoration renovation and professional management of the historic loews jersey theatre . [citation] [citation] these tax abatements have not been without controversy . jersey city residents have criticized these pilot agreements for depriving the public schools of funds and disproportionately saddling owners of non piloted property with local taxes . [citation] republicans from outside of jersey city have criticized fulop for these pilot agreements since piloted buildings do not pay school taxes and thus sustain an artificially high level of state aid . [citation] in jersey city was ranked by atlantic magazine ~~as the th most~~	Small deletion at end of example
Ground truth	he earned a bachelors degree from westminster college in salt lake city . [citation] wheatley was first elected on november and last elected on november . [citation] during the general session he served on the house business and labor committee and the house law enforcement and criminal justice committee . [citation]	Inserted three citations
Full	he earned a bachelors degree from westminster college in salt lake city . wheatley was first elected on november and last elected on november . [citation] during the general session he served on the house business and labor committee and the house law enforcement and criminal justice committee . [citation]	Inserted two of three citations
Baseline	(Predicted no edit)	Predicted no edit
Ground truth	challenge . i am grateful for the efforts their officials have given to this partnersip project . ~~[jpg]~~ baldrige worked during his boyhood as a ranch hand and earned several awards	Deleted image
Full	(Identical to ground truth edit)	Correct edit
Baseline	(Identical to ground truth edit)	Correct edit
Ground truth	brian mast official congressional photo [jpg]~~[png]~~ member of the u .s . house of representatives	Replaced image and inserted text
Full	(Identical to ground truth edit)	Correct edit
Baseline	(Predicted no edit)	Predicted no edit

Model	Edit	Comment
Ground truth	republicans which in turn has led to hecks own ambiguity on supporting trump . [citation] [citation] only days before the election he would not say whether he would or would not vote for trump . [citation] [citation]	Added citation to end of example
Full	republicans which in turn has led to hecks own ambiguity on supporting trump . [citation] [citation] only days before the election he would not say whether he would or would not vote for trump . [citation] He has voted for Trump [citation]	Generated a made-up fact that resolves the ambiguity of the previous sentences; correctly inserted citation.
Baseline	(Predicted no edit)	Predicted no edit
Ground truth	vance ran unsuccessfully for the u .s . senate losing to democratic incumbent patty murray in the election . [citation] in september he announced that he had left the republican party and had	Inserted citation
Full	vance ran unsuccessfully for the u .s . senate losing to democratic incumbent patty murray in the election . he was succeeded by john mccain. [citation] in september he announced that he had left the republican party and had	Generated a made-up fact that is compatible with previous fact; correctly added citation
Baseline	vance ran unsuccessfully for the u .s . senate losing to democratic incumbent patty murray in the election . he was reelected by a margin of to. in. [citation] in september he announced that he had left the republican party and had	Generated a made-up fact that contradicts previous fact; correctly added citation
Ground truth	on january biskupski announced a draft policy on the release of police body camera footage putting her at odds with sim gill the salt lake county district attorney . [citation] [png] in september mayor jackie biskupski announced the formation of a commission against gun violence designed to explore policy questions regarding gun violence and to make funding recommendations to be shared with city county and state officials as well as the salt lake city school district . [citation] on october submitted an ordinance to city council to promote a	Deleted image
Full	(Identical to ground truth edit)	Correct edit
Baseline	on january biskupski announced a draft policy on the release of police body camera footage putting her at odds with sim gill the salt lake county district attorney . [citation] ~~[png] in september mayor jackie biskupski announced the formation of a commission against gun violence designed to explore policy questions regarding gun violence and to make funding recommendations to be shared with city county and state officials as well as the salt lake city school district . [citation]~~ on october submitted an ordinance to city council to promote a	Deleted a large amount of text and four images, including the correct image

Model	Edit	Comment
Ground truth	women steel workers in car~~automobile~~ on picket duty on steel mill property in indiana [jpg] the federal governments inaction permitted state and local authorities and	Changed “automobile” to “car”
Full	~~women~~ steel workers in automobile on picket duty on steel mill property in indiana [jpg] the federal governments inaction permitted state and local authorities and	Incorrectly removed “women”
Baseline	women steel workers ~~in automobile~~ on picket duty on steel mill property in indiana [jpg] the federal governments inaction permitted state and local authorities and	Incorrectly removed phrase “in automobile”
Ground truth	on an old theodore roosevelt adage . [citation] ~~[jpg] davis married the former alvern adams in this historic shreveport house in the highlands neighborhood . it was formerly owned by the eglins the maternal grandparents of davis second successor as governor john j . mckeithen . [citation] [jpg] jimmie davis tabernacle west of quitman ! [jpg] davis grave located in small cemetery behind the tabernacle [jpg] grave of louisiana first lady alvern adams davis who died thirty three years before her husband~~ . davis first wife the former alvern adams the daughter of a physician in	Deleted a large amount of text and four images
Full	(Identical to ground truth edit)	Correct edit
Baseline	(Identical to ground truth edit)	Correct edit
Ground truth	~~biden at the august announcement of her husband becoming barack obamas running mate [jpg]~~ despite personally opposing the iraq war biden had not wanted her husband to	Deleted image and corresponding text
Full	(Identical to ground truth edit)	Correct edit
Baseline	biden at the august announcement of her husband becoming barack obamas running mate [jpg] ~~despite personally opposing the iraq war~~ biden had not wanted her husband to	Incorrectly deleted phrase