FRUIT 🍏: Faithfully Reflecting Updated Information in Text

Robert L. Logan IV[*†1] **Alexandre Passos**[2]
Sameer Singh[1] **Ming-Wei Chang**[†2]
[1]University of California, Irvine [2]Google Research
{rlogan,sameer}@uci.edu
{apassos,mingweichang}@google.com

Abstract

Textual knowledge bases such as Wikipedia require considerable effort to keep up to date and consistent. While automated writing assistants could potentially ease this burden, the problem of suggesting edits grounded in external knowledge has been under-explored. In this paper, we introduce the novel generation task of *faithfully reflecting updated information in text* (FRUIT) where the goal is to update an existing article given new evidence. We release the FRUIT-WIKI dataset, a collection of over 170K distantly supervised data produced from pairs of Wikipedia snapshots, along with our data generation pipeline and a gold evaluation set of 914 instances whose edits are guaranteed to be supported by the evidence. We provide benchmark results for popular generation systems as well as EDIT5—a T5-based approach tailored to editing we introduce that establishes the state of the art. Our analysis shows that developing models that can update articles faithfully requires new capabilities for neural generation models, and opens doors to many new applications.[1]

1 Introduction

Information changes on a constant basis. Every day, athletes are traded to new teams, and musicians and actors produce new albums and TV shows. Maintaining textual knowledge bases to keep track of these changes requires considerable community effort. For instance, a team of 120K volunteer editors make 120 edits to English Wikipedia every minute, and write 600 new articles a day.[2] As the knowledge base grows, the amount of maintenance effort is compounded by the need to keep the knowledge base consistent; e.g., each edit may render information in one of the existing 6.3M+ articles obsolete.

Assistive writing technologies have the potential to substantially reduce the burden of keeping text corpora up to date and consistent. However, existing work has mainly focused on correcting grammar (Wang et al., 2020), reducing repetitive typing (Chen et al., 2019), and following rhetorical directives (Sun et al., 2021), whereas the problem of producing edits grounded in external knowledge has received little attention (Kang et al., 2019). In contrast, numerous works have developed systems for distilling external knowledge into text (e.g., Wikipedia article generation) by treating the problem as multi-document summarization (Liu et al., 2018; Shi et al., 2021) or data-to-text generation (Bao et al., 2018; Parikh et al., 2020). However, these systems are not useful for updating existing texts as they can only generate text from scratch.

To help endow writing assistants with grounded editing capabilities, we introduce the novel generation task of *faithfully reflecting updated information in text* (FRUIT), where the goal is to incorporate new information into an existing piece of text. An illustration is provided in Figure 1. Given an outdated Wikipedia article and collection of new information about the article's subject, FRUIT requires updating the existing text so that it is consistent with the new information, as well as adding text to reflect new salient facts, e.g., in Figure 1, the first sentence is updated to reflect that Tom Kristensson now drives in the Junior World Championship, and new sentences are added to reflect his achievements in 2019 and 2020.

FRUIT presents several unique challenges. First, unlike many generation tasks, models cannot obtain good performance by solely relying on their parametric world knowledge. Whenever the provided evidence contradicts parametric knowledge, the model must prefer the evidence, which recent work has shown is difficult for pretrained language

*Work done during an internship at Google Research.
†Corresponding Author.
[1]Our data and code are available at: https://github.com/google-research/language/tree/master/language/fruit.
[2]https://en.wikipedia.org/wiki/Wikipedia:Statistics

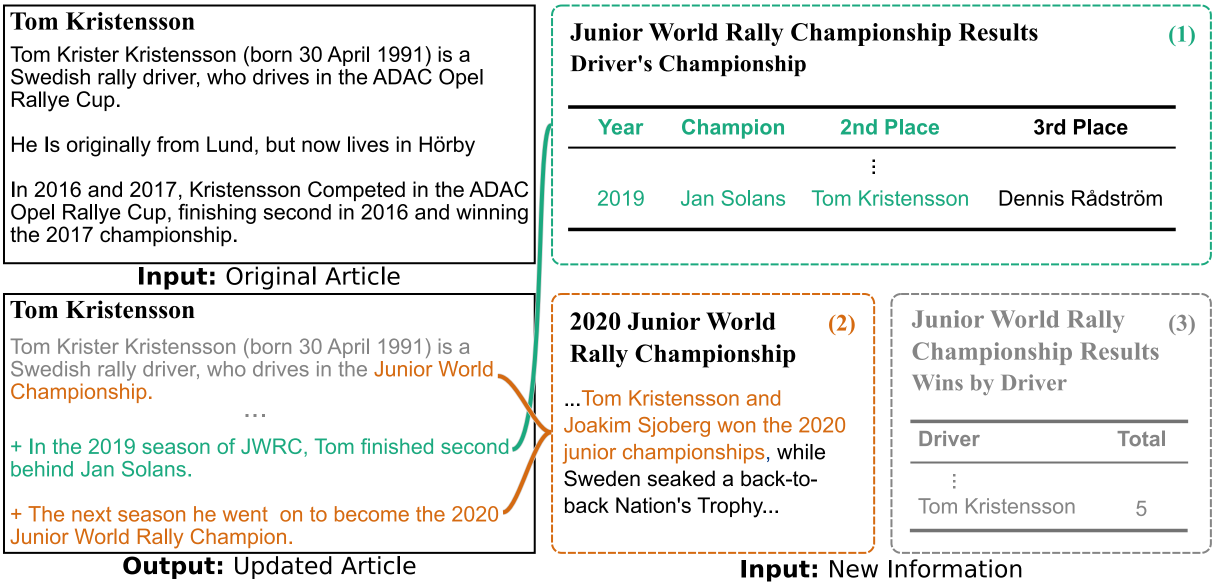

Figure 1: **Illustration of the FRUIT task**. An outdated *original article* and relevant *new information* are provided as inputs, and the goal is to generate the *updated article*. In this example, the original article about Tom Kristensson was written in 2020, and the new information is comprised of updated information about Tom Kristensson that has been added to other Wikipedia articles between 2020 and 2021. Given these inputs, the goal is to produce the updated 2021 version of article. Models need to identify the relevant supporting facts (orange and teal) to generate faithful updates while ignoring superfluous information (grey).

models (Krishna et al., 2021; Longpre et al., 2021). Second, the generated text needs to be faithful to *both* the original article and the new evidence, *except* when evidence invalidates information in the existing article. Finally, this task requires models to jointly read and analyze evidence from both textual and tabular sources and determine which is relevant and which can be ignored, thus combining challenging aspects of both multi-document summarization and data-to-text generation.

To facilitate research on this task, we release the FRUIT-WIKI dataset, a collection of over 170K distantly supervised ("*silver*") update-evidence pairs. This dataset is produced by comparing pairs of English Wikipedia snapshots to identify updates to an article between two snapshots, and associating information from the other articles that supports these updates under a distant supervision assumption. As there is no guarantee that updates in the later Wikipedia snapshots can be supported by the collected evidence, we also collect a "*gold*" evaluation set of 914 human annotated update-evidence pairs where unsupported claims have been removed without disturbing fluency. We train and validate our models using silver data and then evaluate the final performance using gold data.

We establish initial benchmark results for a number of trivial and neural sequence-to-sequence baselines. We also introduce EDIT5, a T5-based model specially adapted for grounded editing, which establishes state-of-the-art performance on FRUIT-WIKI. Through an extensive set of analyses, we identify a number of failure modes needed to be improved upon in order to obtain better performance on FRUIT-WIKI, as well as other interesting topics for future work on this task. We additionally release our data collection pipeline to allow researchers to produce data from future Wikipedia snapshots and other languages, which we show to produce high-quality silver data.

2 The FRUIT Task

2.1 Task Definition

In this section we introduce the task of *faithfully reflecting updated information in text* (FRUIT). Given an input piece of text focused on a topic or event, along with a collection of potentially new information about the subject of the text, the goal is to update the input text to reflect the new information. A concrete illustration of the task is provided in Figure 1. The original piece of text along with its updates are shown on the left, while the new information is shown on the right.

Formally, we assume access to pair of texts, A^t and $A^{t'}$, pertaining to a given subject, written at times t and t' (respectively). In addition, we assume access to a set of new information, a.k.a., evidence, $\mathcal{E}^{t \to t'} = \{E_1, \ldots, E_{|\mathcal{E}|}\}$, mentioning the subject written between times t and t'. As is shown in Figure 1, the evidence can contain structured objects (e.g., excerpts from tables) as well as unstructured text. Given A^t and $\mathcal{E}^{t \to t'}$ the goal is produce the updated text $A^{t'}$.

Successful completion of this task requires a number of complex and inter-related reasoning capabilities. For one, models must be able to identify which evidence contradicts existing portions of the source article, and which evidence introduces new salient information about the subject in order to correctly choose whether to alter the existing text vs. add new text. For example, in Figure 1 the first sentence is updated to reflect that Tom Kristenssson now races in a different competition, whereas new sentences are added describing his achievements in the years 2019 and 2020. Models must also be able to determine whether a given piece of evidence should be used at all, i.e., perform content selection. For example, in Figure 1, the number of rounds won by Kristennsen appears in the evidence but does not correspond to any piece of updated text. Although some evidence may not appear in the updated article, the converse is not true, the system should aim to generate an updated article where all the updates are faithful to the evidence.

2.2 Evaluation

In this section we introduce important considerations for evaluating FRUIT systems.

Evaluate on Updated Text There is often considerable overlap between the original and updated text. As we will see in Section 5 this poses a challenge for standard evaluation metrics like ROUGE (Lin, 2004) as systems can achieve high scores without making any updates. In this work, we propose to evaluate FRUIT systems using an alternative metric, UpdateROUGE, that only considers updated sentences instead full texts. For example, in Figure 1, the reference for UpdateROUGE only consists of the first and last two sentences.

Evaluate Faithfulness Ensuring that generations faithfully reflect information in the evidence and updated article is crucial. However measuring faithfulness of generations is an active area of research (Çelikyilmaz et al., 2020) and adapting existing metrics to the FRUIT task is non-trivial.

As a simple proxy for faithfulness, we choose to measure the token overlap between named entities appearing in the generation and the target article/evidence, where entities are identified using the named entity recognizer used by Guu et al. (2020) to perform salient span masking. We specifically introduce the following measurements:

1. **Unsupported Entity Tokens.** This metric shows the average number of entity tokens appearing in generated updates that do not appear in the *source article* or *evidence*. This is intended to capture the overall amount of unfaithful text, focusing on entities, where higher numbers indicate less faithfulness.
2. **Entity Precision and Recall.** These metrics capture the overlap of mentions between the generated updates and the *target*. Entity precision measures the fraction of entity tokens appearing in the generated updates that appear in target entities, whereas entity recall measures the fraction of entity tokens in the target that appear in the entities in generated updates. The latter is similar to UpdateROUGE but only evaluated on entities, and thus, potentially less sensitive to paraphrasing.

Parametric Knowledge Consideration FRUIT systems should incorporate information from the provided evidence into the update, and not information that happened to be present during training or pretraining. In this work we attempt to address this by evaluating models only on updates that were made to the text after the data used to pretrain and finetune the model was collected. As this setup precludes evaluating models trained after 2020 on FRUIT-WIKI, we release our data collection pipeline so that researchers can produce evaluation datasets from future versions of Wikipedia.

3 Dataset Collection and Analysis

As discussed in the introduction, keeping track of new information and then updating articles to reflect that information requires a massive amount of manual effort. Thus, in order to scalably collect sufficient data for training and evaluating FRUIT systems, some amount of automation is likely required. In this section we introduce the FRUIT-WIKI dataset and associated data collection pipeline, which allows the automatic collection of high-quality training and evaluation data for FRUIT from pairs of Wikipedia snapshots.

	Train	Test	
		Silver	Gold
Years	'19-'20	'20-'21	'20-'21
Articles	114K	54K	914
Edits	407K	182K	3.0K
Subst. Edits	135K	62K	1.3K
Evidence	720K	315K	7.7K
Content Sel.	93K	42K	913

Table 1: **Dataset Statistics.** We use 10% of the training data as our validation data.

UpdateROUGE			Entity	
1	2	L	Prec.	Recall
87.4	84.6	87.1	91.8	94.6

Table 2: **Inter-Annotator Agreement**.

3.1 Pipeline

Our data collection pipeline produces distantly annotated training and evaluation data from pairs of Wikipedia snapshots. We will refer to the earlier snapshot as the *source* snapshot, and the later snapshot as the *target* snapshot.

Step 1. Collect Article Updates We compute the diff between the introductory sections of articles appearing in both the *source* and *target* snapshot to identify all of the material that has been updated (which will serve as A^t and $A^{t'}$). We also compute the diff between the non-introductory sections of articles to find new mentions of the subjects of other articles (which will serve as $\mathcal{E}^{t \to t'}$). These mentions can take the form of sentences in the text, as well as new table rows and list entries. Entities are disambiguated using Wikipedia hyperlinks.

Step 2. Filter Stylistic Updates A large number of edits to Wikipedia are stylistic (Daxenberger and Gurevych, 2012), and are therefore irrelevant to our task. In the next step of the pipeline, we attempt to filter articles that have only been superficially edited by keeping only those where at least one new *added entity* appears in the *target* snapshot.

Step 3. Identify Supporting Evidence In the last step of our pipeline, we seek to determine which pieces of evidence in $\mathcal{E}^{t \to t'}$ justify each of the updated sentences in $A^{t'}$. To do so, we make the following distant supervision assumption: an updated sentence $a \in A^{t'}$ containing an *added entity* s' is substantiated by a piece of evidence $E \in \mathcal{E}^{t \to t'}$ only if s' is also mentioned in E. The accuracy of the annotations produced by this assumption will be measured in Section 3.3.

Our pipeline is implemented using Apache Beam,[3] to allow for distributed processing. We plan on releasing the code upon publication to enable other users to produce FRUIT data from future Wikipedia snapshots, as well as languages other than English.

3.2 FRUIT-Wiki

We run our pipeline on English Wikipedia snapshots from Nov. 20, 2019 to Nov. 20, 2020 to produce the training dataset, and from Nov. 20, 2020 to June 1, 2021 to produce the evaluation dataset. Detailed statistics are provided in Table 1. On average, there are around 3 to 4 updates per article, and around 7 pieces of associated evidence. About 80% of updates require some form of content selection, i.e., ignoring some evidence, when performing updates.

We find that only a third of the updates are substantiated by one or more pieces of evidence according to our distant supervision assumption. Thus, the remaining updates are either: a) superficial changes to the source article, or b) additions of new claims that are unsupported with respect to the collected evidence. The latter is a particular issue as these claims can cause the model to learn to hallucinate during training, and should be impossible for the model to guess during evaluation. Through the usage of human annotations and carefully selected evaluation metrics we will study the extent to which this is an issue throughout the rest of the paper.

We categorize articles in our dataset using the Wikimedia Foundation's topic model (Asthana and Halfaker, 2018). The distribution of topics is displayed in Figure 2. We find that the majority (approximately 50%) of updates deal with cultural topics (e.g., sports, media, personal biographies), and geographic entities (e.g., countries, states) which intuitively are likely to be affected by current events. while there are few updates to STEM- and history-related articles.

3.3 Gold Evaluation Data

To address the issue of unsupported claims during evaluation, we hired a team of 9 annotators to produce a *"gold"* evaluation subset of our test dataset.

[3] https://beam.apache.org/

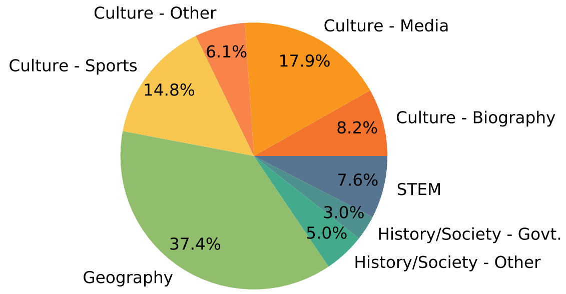

Figure 2: **Topic Distribution.**

We collect annotations for 914 update-evidence pairs where each instance is corrected to ensure that all of the updates are supported. For the remainder of the paper we will refer to the distantly supervised test dataset annotations as *"silver"*.

Annotation Process For each instance, annotators were shown the source article, evidence, and a marked up copy of the target article. In the marked up article, each updated sentence was highlighted and prefixed with reference labels to the supporting evidence identified by our pipeline. The correction process proceeded in two steps. In the first step, annotators were asked to highlight all of the unsupported claims and incorrect reference labels in the target article. In the second step, annotators were then asked to remove the unsupported text and minimally update the article to preserve fluency.

Annotators attended an initial 30 minute training and were provided regular feedback from the authors during the early stages of annotation. To ensure data quality, an additional annotator was hired with the sole job of checking the other annotator's work and correcting their mistakes. In total annotators spent roughly 500 hours on annotation. The annotation interface and a completed annotation are shown in Figure A7 in the Appendix.

Agreement We measure annotator agreement using a subset of 100 instances that were annotated by multiple annotators. Following Chen et al. (2015) and Shi et al. (2021), we quantify agreement by computing the evaluation metrics described in Section 2.2. The results are provided in Table 2. We observe high inter-annotator agreement with all scores in the 80s and 90s.

Analysis Statistics for the gold evaluation dataset are provided in Table 1. Overall, they closely resemble the statistics for the distantly supervised data with one exception: the fraction of substantiated updates has increased.

UpdateROUGE			Entity		Reference Agreement
1	2	L	Prec.	Recall	
83.7	81.2	83.4	90.4	100.0	84.5

Table 3: **Gold and Silver Annotation Agreement**. Quality of Silver Annotations by using the Gold as the reference.

To measure the quality of our silver data, we re-apply the approach used to measure inter-annotator agreement to compute agreement between the gold and silver annotations. We also measure the *reference agreement*, i.e., the fraction of reference labels kept by the annotators. Results are provided in Table 3. We find that agreement is high with most scores in the 80s, a strong indication that the data produced by our pipeline is high quality. In particular, the high UpdateROUGE scores provide further evidence that only a small amount of the updated text in the weakly supervised data is unsupported, while the high reference agreement indicates that our distant supervision assumption is usually accurate.

4 Methods

In this section we introduce baseline methods to establish initial benchmark results on FRUIT-WIKI. We consider trivial approaches that copy task inputs, as well as T5, a neural sequence-to-sequence baseline which has shown strong performance on related tasks such as summarization (Raffel et al., 2020; Rothe et al., 2021) We additionally introduce EDIT5, a variant of T5 that produces a sequence of edits instead of the entire updated text, and employs additional tweaks to improve performance.

4.1 Copy Baselines

The first set of baselines we introduce are trivial methods that merely copy the input. We consider two variants:

- **Copy Source**: Generates a copy of the source article, and
- **Copy Source + Evidence**: Generates a copy of the source article concatenated with the evidence.

Our evaluation metrics only apply to unstructured text, however the evidence may contain structured tables. In order to convert these tables to text, we apply a conventional linearization scheme (Lebret

	UpdateROUGE			Entity		Unsup.
	1	2	L	Prec.	Recall	Tokens
Copy Source	0.0	0.0	0.0	0.0	0.0	0.00
+ All Evidence	18.8	6.9	12.0	37.9	64.9*	0.00
T5-Large	31.1	18.4	24.4	52.7	44.9	2.67
+ Evidence Input	44.3	29.4	36.8	62.2	50.7	2.34
EDIT5-Small	41.2	27.3	35.3	62.4	44.9	1.71
EDIT5-Base	47.0	32.1	39.7	62.2	**54.9**	2.28
EDIT5-Large	46.3	32.4	39.6	67.2	53.1	**1.54**
EDIT5-3B	**47.4**	**34.0**	**41.1**	**69.9**	52.5	1.58

*Entity recall is not 100% for the Copy Source + All Evidence baseline due to lexical variation in entity mentions.

(a)

Grounded Updates	50
Additional Content	15
Missing Content	22
Ungrounded Updates	35
Number/Date	21
Distorted Evidence	11
Hallucination	14
No Updates	14

(b)

Table 4: **(a) Model Results on Gold Evaluation Data.** EDIT5 outperforms T5 models in all metrics. **(b) Error Analysis for EDIT5-3B**. We find that the model makes correct, grounded updates on 50% of the inspected articles. For incorrect updates, ungrounded numbers/dates are one of the main sources of error.

et al., 2016; Wiseman et al., 2017) that separates table entries using row and column delimiters.

4.2 T5

T5 (Raffel et al., 2020) is a pretrained sequence-to-sequence (Sutskever et al., 2014) model based on the transformer architecture (Vaswani et al., 2017). Similar to the previous section we experiment with two variants:

- **T5**: Only includes the source article in its input,
- **T5 + Evidence Inputs**: Includes both the source article and evidence in the input.

Tabular inputs are linearized using the same approach described in the previous section. Experiments are performed using the JAX-based T5X library.[4] Hyperparameters and additional training details are described in Appendix B.

4.3 EDIT5

Lastly, we introduce EDIT5, which improves upon the T5-based approach described in the previous section through the usage of a compressed output format that removes the need to write the entire update from scratch and encourages content planning. The output is modified in two ways:

First, as the majority of text in the target article is copied from the source, we replace any copied sentence with a single *copy token* identifying the sentence, e.g., if the second sentence is copied it is replaced by the token `[2]`. Similar to a copy mechanism (See et al., 2017), this allows the model to dedicate less capacity to repeating sequences from the input. As the resulting output resembles that produced by the `diff` data comparison utility, we refer to this as a diff-formatted output.

[4]https://github.com/google-research/t5x

```
(2) Tom Krister Kristensson (born
30 April 1991) is a Swedish rally
driver, who drives in the Junior World
Championship.  [1] [2] (1) In the 2019
season of JWRC, Tom finished second
behind Jan Solans.  (2) The next season
he went on to become the 2020 Junior
World Rally champion.
```

Figure 3: **EDIT5 Output Format.** Instead of generating the fully updated text, EDIT5 generates sequences of edited sentences, copy tokens (e.g., `[2]`, which means copy the second sentence), and reference tokens (e.g., `(1)`, which means the following sentence should use the first piece of evidence).

Second, before each update we insert a sequence of *reference tokens* identifying the pieces of evidence that support the update, e.g., if the first and third piece of evidence in $\mathcal{E}^{t \to t'}$ support an update then the update is prefaced by `(1)(3)`. This approach, inspired by the use of entity chains for summarization (Narayan et al., 2021), trains the model to plan which references to use before generating an update. These reference tokens are removed from the output text of the model prior to computing the evaluation metrics.

An example of the EDIT5 output format is provided in Figure 3, and a comparison to the T5 output format is provided in Appendix D. Training details and hyperparameters match the setup described in Section 4.2.

5 Results and Analysis

Baseline results on the gold evaluation data are provided in Table 4a, and ablation results are provided in Appendix A. In general, we find that the copy baselines perform worse than T5 and T5 performs worse than EDIT5. Notably, the copy source baseline rightfully scores zero on all metrics, while we will later find that it obtains a high ROUGE score.

Although our models are trained on silver data, they still obtain good performance on the gold evaluation set. This shows the high quality of our silver data collection pipeline, and T5's ability to generate reasonable updates based on the evidence.

For the T5 baselines, we find that adding evidence to the input results significant increase in all metrics, demonstrating that using the evidence is crucial to obtaining good performance.

EDIT5 obtains additional 3-5% absolute increase in all performance metrics compared to T5, establishing EDIT5 as a strong baseline for future systems to be compared against. The reduction of unsupported entity tokens implies that EDIT5 hallucinates less frequently than T5 models. Results are provided for different model sizes to illustrate how performance scales with parameter counts.

Example Output An example EDIT5 output is provided in Figure 4, and additional outputs in Appendix E. The examples illustrate important features of the task. In Figure 4 the goal is to update the Wikipedia article for Holli Sullivan to reflect her new role of Secretary of State of Indiana. In the reference, this information is reflected in an updated version of the first sentence as well as in a newly added last sentence. An additional sentence is added after the first sentence paraphrasing the introduction of the source article, which describes Sullivan's previous position as a member of the Indiana House of Representatives.

In the EDIT5 output for this example, information is only added at the end of the article. While the model correctly states that Sullivan was appointed to be Secretary of State by Governor Eric Holcomb, as well as includes additional context surrounding Sullivan's appointment that is paraphrased from the evidence, there are some issues with the output. First, because the first sentence of the article is not updated there is conflicting information about Sullivan's current position. Second, the added sentence hallucinates that Sullivan was appointed in January 2020 when she was actually appointed in March 2021, a fact that directly appears in the evidence.

	ROUGE		
	1	2	L
Copy Source	78.1	69.3	75.0
T5-Large	57.0	44.2	49.5
EDIT5-Large	78.6	69.1	72.7

Table 5: **ROUGE Scores Are Insensitive to Edits.**

Categorizing Errors To better understand the types of errors made by EDIT5, we review a random sample of 100 of its predictions on the gold evaluation data and categorize them as either: *grounded updates*, meaning all generated claims are supported, *ungrounded updates*, meaning at least one unsupported claim appears in the output, or *no updates*, meaning the model did not predict any updates. For grounded updates we additionally keep track of how many updates include *additional content* not present in the ground truth update, or are *missing content* that appears in the ground truth update. For ungrounded updates we track whether an incorrect *number/date* appears in the update, the model *distorted evidence*, i.e., paraphrased or combined claims in the evidence in a way that changed their meaning, or *hallucinated* new claims unrelated to the evidence.

The results of this analysis are presented in Table 4b. We find that EDIT5 makes no mistakes on half of the examples, however a substantial portion of these updates had some issue with content selection. Of the incorrect updates, the most common mistake was incorrect numbers and dates, followed by hallucinations, and finally distorted evidence. This suggests that improving numeracy could be a fruitful line of study in future work on this task.

ROUGE is Problematic We provide ROUGE F-scores for each of the baseline models on the gold evaluation data in Table 5. In contrast to the previous results, we find that the simple copy source baseline attains a strong score of 77.4 despite making no updates. This is better than the T5 baseline results and comparable to the EDIT5 results. This illustrates the importance of evaluating on updates rather than the whole text.

Silver Data is Useful for Evaluation The results in Section 3.3 demonstrate high agreement between the silver and gold evaluation data which begs the question: can silver data be used in place

Original Article

Holli Sullivan is an American politician who serves in the Indiana House of Representatives as a member of the Republican Party. In 2014 the district 78 seat for state Representative was vacated by Suzanne Crouch, who had been appointed state Auditor. …Text omitted to save space… In 2017, she co-authored House Bill 1002, which provided for a long term plan for sustaining roads and bridges in Indiana including a phase-in shift of all gas tax to be dedicated to a dedicated infrastructure fund. That same session she authored a bill which created a strategic plan to reduce cervical cancer.

New Information

Secretary of State of Indiana
List of Secretaries of State

#	Name	Took Office	Left Office
62	Holli Sullivan	March 16, 2021	-

Secretary of State of Indiana
Introduction
The current office holder is Holli Sullivan, who was appointed by Governor Eric Holcomb to serve out the term of former Secretary of State Connie Lawson, who announced in February 2021 that she planned on resigning from office.

Ground Truth

Holli Sullivan is an American politician who is the 62nd and current secretary of state of Indiana since March 2021. As a member of the Republican Party, she previously represented the 78th district in the Indiana House of Representatives from 2014 to 2021. …Copied text… In 2021, Holli was named the 62nd Secretary of State of Indiana by Governor Eric Holcomb.

EDIT5

Copied text... In January 2020 Representative Sullivan was appointed by Governor Eric Holcomb to serve out the term of former Secretary of State Connie Lawson, who announced in February 2021 that she planned on resigning from office.

Figure 4: **Example EDIT5 Output vs Ground Truth.** Color coding indicates alignment between the new information and the edits. EDIT5 updates the original article by paraphrasing sentences from the textual evidence, however misses relevant information in the table, and generates an incorrect date.

UpdateROUGE			Entity		Unsup.
1	2	L	Prec.	Rec.	tokens
100.0	100.0	94.3	75.4	92.8	92.8

Table 6: **Spearman Rank Correlation Between Gold and Silver Performance Metrics**.

	UpdateROUGE			Entity		Unsup.
	1	2	L	Prec.	Rec.	Tokens
EDIT5	46.3	32.4	39.6	67.2	53.1	**1.54**
Control	**57.6**	**42.1**	**50.2**	**70.5**	**64.5**	2.42

Table 7: **Controllability**. Using control codes that indicate which sentences to delete, add or edit, and which evidence to use, can greatly improve generation.

of gold data for evaluation? To answer this, we measure the Spearman rank correlation between the gold baseline results in Table 4a and silver baseline results (provided in Table A2 of the Appendix to save space). Rank correlations for each of the metrics are shown in Table 6. Overall we find high rank correlation for each of the metrics, which suggests silver evaluation performance is a reliable indicator of gold performance. Thus, models whose pretraining data overlaps FRUIT-WIKI may be evaluated and compared on data produced by running our pipeline on future Wikipedia snapshots without requiring further human evaluation.

Controllability The improvement we obtained from EDIT5 over T5 implies that more controls can be added into the model. In this section we investigate whether additional control provided by the users can improve the overall generations. We follow Keskar et al. (2019) and Narayan et al. (2021), and provide more detailed instruction by adding *control codes*, i.e., special tokens, to the *input* that instruct the model whether to add, copy, edit or remove a sentence, as well as which evidence to use when making an addition or edit. We use the target text to provide oracle labels for the control code, and see if the EDIT5 can take advantage of the codes. Example inputs and predictions are provided in Figure A6 of the Appendix.

Results on the gold evaluation data are provided in Table 7. Including oracle control codes in the input produces a substantial 10% absolute improvement in all metrics besides unsupported tokens. This demonstrates that increased user control has the potential to produce updates that more closely resemble the desired output.

6 Related Work

Early work on writing assistants largely focuses on grammar error correction; for a survey see Wang et al. (2020). Neural models have expanded the capabilities of writing assistants to solve a wider variety of tasks including: autocompletion (Chen et al., 2019), and following rhetorical directives such as paraphrasing, elaborating, etc. (Sun et al., 2021). In this work, we seek to expand these capabilities further to producing grounded updates, which has been previously studied by Kang et al. (2019), however only for post-modifier generation.

As our primary focus is on writing grounded updates to Wikipedia articles, our work is closely related to existing works on Wikipedia article generation, which generally uses one of two approaches: data-to-text generation (Lebret et al., 2016; Bao et al., 2018; Parikh et al., 2020; Chen et al., 2021; Cheng et al., 2020), or multi-document summarization (Banerjee and Mitra, 2016; Liu et al., 2018; Shi et al., 2021). In particular, the hyperlink-based approach for associating evidence to articles is directly inspired by these works, and our annotation procedure for removing unsupported text directly draws from Parikh et al. (2020).

Determining which facts contradict claims in the existing article is a central topic of work on fact extraction and verification (Thorne et al., 2018). Recently, Schuster et al. (2021) introduced the VITAMIN-C dataset of factual revisions to Wikipedia articles and the task of factually consistent generation. This work differs from FRUIT in that it only focuses on sentences and does not require adding new facts or content selection.

Our work is also related to the TAC 2008 Update Summarization Task (Dang and Owczarzak, 2008), which involves summarizing information about a topic that does not overlap with an existing summary, instead of updating an existing summary to reflect new information.

7 Conclusion and Future Work

In this work we introduced FRUIT, a novel text generation task where the goal is to update an article to reflect new information about its subject. To enable research on this task, we formulated a pipeline for extracting weakly supervised training and evaluation data from pairs of Wikipedia snapshots, and collected data for the years 2019-2020 and 2020-2021, as well as human annotated gold evaluation data. We additionally provided results for several strong baselines, that demonstrate both the feasibility of this task, as well as strong correlation between gold and distantly supervised data evaluation performance that establishes the trustworthiness of future data produced using our pipeline.

This work lays the foundation for future research into making faithful updates to entries in textual knowledge bases. One limitation of this work is that the metrics we use to evaluate faithfulness are all entity-centric, and thus may overstate the performance of models whose edits include the correct entities but misspecify the relations between them or other facets of the evidence. Accordingly, one promising direction for future work on this task is to develop more robust metrics for measuring faithfulness.

An additional promising direction for future work is to consider open settings of evidence collection, where other forms of updated information such as excerpts from news articles could be used to justify edits in place of updates to other entries in the same knowledge base. Relatedly, we also recommend studying this task in streaming settings, where updates arrive in sequential fashion, in addition to the batch setting considered in this work.

Finally, there are a number of promising directions for improving model performance on this task. In particular, copy mechanisms (See et al., 2017) have been widely used in data-to-text tasks (Wiseman et al., 2017), and may help mitigate issues such as the mistranscribed dates we saw in Section 5.

Acknowledgements

We would like to thank Kelvin Guu, Kenton Lee, Kristina Toutanova, and the anonymous reviewers for their helpful feedback. This work is funded in part by the DARPA MCS program under Contract No. N660011924033. Additionally, Robert is supported in part by the Irvine Initiative in AI, Law, and Society fellowship.

Ethical Considerations

This paper introduces a dataset and system for updating an existing piece of text to incorporate information from external evidence. Depending on the veracity of the external evidence, systems for solving this task could potentially be abused by bad actors to spread misinformation.

References

Sumit Asthana and Aaron Halfaker. 2018. With few eyes, all hoaxes are deep. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Siddhartha Banerjee and Prasenjit Mitra. 2016. Wikiwrite: Generating wikipedia articles automatically. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2740–2746. IJCAI/AAAI Press.

Junwei Bao, Duyu Tang, Nan Duan, Zhao Yan, Yuanhua Lv, Ming Zhou, and Tiejun Zhao. 2018. Table-to-text: Describing table region with natural language. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5020–5027. AAAI Press.

Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail smart compose: Real-time assisted writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 2287–2295. ACM.

Mingda Chen, Sam Wiseman, and Kevin Gimpel. 2021. WikiTableT: A large-scale data-to-text dataset for generating Wikipedia article sections. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 193–209, Online. Association for Computational Linguistics.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *ArXiv*, abs/1504.00325.

Liying Cheng, Dekun Wu, Lidong Bing, Yan Zhang, Zhanming Jie, Wei Lu, and Luo Si. 2020. ENT-DESC: Entity description generation by exploring knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1187–1197, Online. Association for Computational Linguistics.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the tac 2008 update summarization task. *Theory and Applications of Categories*.

Johannes Daxenberger and Iryna Gurevych. 2012. A corpus-based study of edit categories in featured and non-featured Wikipedia articles. In *Proceedings of COLING 2012*, pages 711–726, Mumbai, India. The COLING 2012 Organizing Committee.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training.

Jun Seok Kang, Robert Logan, Zewei Chu, Yang Chen, Dheeru Dua, Kevin Gimpel, Sameer Singh, and Niranjan Balasubramanian. 2019. PoMo: Generating entity-specific post-modifiers in context. 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 826–838, Minneapolis, Minnesota. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *ArXiv*, abs/1909.05858.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. 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.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonccalo Simoes, and Ryan T. McDonald. 2021. Planning with entity chains for abstractive summarization. *ArXiv*, abs/2104.07606.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and

Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Sascha Rothe, Joshua Maynez, and Shashi Narayan. 2021. A thorough evaluation of task-specific pretraining for summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 140–145, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR.

Weijia Shi, Mandar Joshi, and Luke Zettlemoyer. 2021. DESCGEN: A distantly supervised datasetfor generating entity descriptions. 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 415–427, Online. Association for Computational Linguistics.

Simeng Sun, Wenlong Zhao, Varun Manjunatha, Rajiv Jain, Vlad Morariu, Franck Dernoncourt, Balaji Vasan Srinivasan, and Mohit Iyyer. 2021. IGA: An intent-guided authoring assistant. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5972–5985, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yu Wang, Yuelin Wang, Jie Liu, and Zhuo Liu. 2020. A comprehensive survey of grammar error correction. *ArXiv*, abs/2005.06600.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Asli Çelikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *ArXiv*, abs/2006.14799.

Appendix

A Ablation Study

We perform an ablation study to measure the impact of the modifications made to the target output of EDIT5. The results are provided in Table A1 We observe that both the diff format and including reference tokens have a positive impact on the evaluation metrics, with reference tokens having the larger impact.

	UpdateROUGE			Entity		Unsupp.
	1	2	L	Prec.	Rec.	Tokens
EDIT5	**46.3**	**32.4**	**39.6**	**67.2**	**53.1**	**1.54**
- Diff	45.5	31.7	39.1	66.8	50.8	1.66
- Ref.	45.1	31.6	38.8	66.3	50.7	1.89

Table A1: **EDIT5 Ablations**.

B Model Training Details

Optimizer: AdaFactor (Shazeer and Stern, 2018), Batch Size: 128, Learning Rate: 1e-3, Dropout Rate: 0.1, Training Iterations: 30,000. Training performed on a cluster of 16 2nd generation TPUs for <3B param models, and 32 TPUS for 3B parameter models.

C Silver Baseline Results

	UpdateROUGE			Target Entity		Evid.
	1	2	L	P	R	Acc
T5-Large	26.8	15.9	22.3	56.3	29.8	2.33
+ Evid.	39.2	27.3	34.2	66.9	42.4	1.63
EDIT5						
Small	37.8	24.9	32.6	61.4	41.2	1.53
Base	42.8	28.7	36.4	60.5	**49.2**	2.32
Large	42.7	29.9	37.2	66.1	47.5	1.47
3B	**43.8**	**31.5**	**38.6**	**68.4**	48.6	1.53

Table A2: **Baseline Results on Silver Evaluation Data.**

D Input and Output Formats

```
(2) [0] Elizabeth Lynne Cheney (; born July 28, 1966) is an American attorney
and politician serving as the U.S. Representative for since 2017.  [1] Cheney is
the House Republican Conference Chair, the third-highest position in GOP House
leadership.  [2] She is the third woman elected to that position after Deborah
Pryce and Cathy McMorris Rodgers.  [3] Cheney is the elder daughter of former
Vice President Dick Cheney and Lynne Cheney.  [4] She held several positions in
the U.S. State Department during the George W. Bush administration.  [5] She has
been politically active on behalf of the Republican Party and is a co-founder
of Keep America Safe, a nonprofit organization concerned with national security
issues.  [6] She was a candidate for the 2014 election to the United States Senate
in Wyoming, challenging the three-term incumbent Mike Enzi, before withdrawing
from the race.  [7] In the House of Representatives, she holds the seat that was
held by her father from 1979 to 1989.  [8] She is known for her hawkish foreign
policy views.  [CONTEXT] (0) Andy Biggs U.S. House of Representatives - Tenure -
2021 storming of the United States Capitol On January 12, 2021, Biggs called on
fellow GOP Representative Liz Cheney (R-WY) to resign from her leadership position
within the Republican Caucus, after she voted in favor of Donald Trump's second
impeachment.  (1) 116th United States Congress Leadership - House of Representatives
- Minority (Republican) leadership * House Minority Leader and Chair of the House
Republican Steering Committee:  Kevin McCarthy * House Minority Whip:  Steve Scalise
* Chair of the House Republican Conference:  Liz Cheney * Vice Chair of the House
Republican Conference:  Mark Walker * Secretary of the House Republican Conference:
Jason Smith * Chair of the House Republican Policy Committee:  Gary Palmer * Chair
of the National Republican Congressional Committee:  Tom Emmer * House Republican
Chief Deputy Whip:  Drew Ferguson (2) A Call for American Renewal INTRODUCTION The
manifesto was released one day after the ousting of Representative Liz Cheney as
chair of the House Republican Conference, and was largely seen as a reaction against
the influence of Trumpism within the Republican Party.  (3) List of nicknames used
by Donald Trump Domestic political figures - Table-0-11 [HEADER] [COL] Nickname
[COL] Personal name [COL] Notes [ROW] id="The Warmonger" [COL] The Warmonger [COL]
Liz Cheney [COL] U.S. representative for Wyoming's at-large congressional district;
Chair of the House Republican Conference (4) Conscience vote Practice in various
countries - United States Similarly, when House Republican leadership decided
not to whip votes against the second impeachment of Donald Trump, Liz Cheney--the
third-highest-ranking Republican--referred to the matter as a "vote of conscience".
```

Figure A1: **Input Format.**

```
(2) Elizabeth Lynne Cheney ( ; born July 28, 1966) is an American attorney and
politician who has served as the U.S. Representative for since 2017.  She was
the Chair of the House Republican Conference, the third-highest position in the
House Republican leadership.  She is the third woman elected to that position
after Deborah Pryce and Cathy McMorris Rodgers.  She held several positions in
the U.S. State Department during the George W. Bush administration, notably as
Deputy Assistant Secretary of State for Near Eastern Affairs and Coordinator for
Broader Middle East and North Africa Initiatives.  She promoted regime change in
Iran while chairing the Iran Syria Policy and Operations Group with Elliott Abrams.
In 2009 Cheney and Bill Kristol founded Keep America Safe, a nonprofit organization
concerned with national security issues that advocated the positions of the former
Bush administration.  She was a candidate for the 2014 election to the U.S. Senate
in Wyoming, challenging three-term incumbent Mike Enzi, before withdrawing from the
race.  In the House of Representatives, she holds the seat her father held for a
decade, representing Wyoming from 1979 to 1989.  Cheney is a neoconservative.  She
later supported the second impeachment of Donald Trump for his role in the 2021
storming of the U.S. Capitol.
```

Figure A2: **T5 Output Format.**

```
(2) [0] [1] [2] [3] [4] [5] [6] In the House of Representatives, she holds the
seat that was held by her father from 1979 to 1989.  (6) She is known for her
neoconservative foreign policy views, and her affiliation with the Trump campaign.
(0) (1) (2) (3) (4) Cheney is under fire for her role in the second impeachment of
Donald Trump in January 2021.
```

Figure A3: **EDITT5 Output Format.**

E More Qualitative Examples

Original Article

Mike McMeeken (born 10 May 1994) is an English rugby league footballer who plays as a forward for the Castleford Tigers in the Super League. McMeeken has also represented England at international level, playing in two games at the 2017 World Cup. He started his career in the Super League with the London Broncos, also playing on loan in League 1 at the London Skolars before joining the Tigers.

New Information

Castleford Tigers 2021
Transfers - Losses

Player	Club	Contract	Date
Mike McMeeken	Catalans Dragons	2 Year	December 2020

Catalans Dragons 2021
Transfers - Gains

Player	Club	Contract	Date
Mike McMeeken	Castleford Tigers	3 Year	June 2020

Ground Truth

Mike McMeeken (born 10 May 1994) is an English rugby league footballer who plays as a forward for the Catalans Dragons in the Super League…Copied text...He joined Catalans Dragons in December 2020, ahead of the 2021 season.

EDIT5

Mike McMeeken (born 10 May 1994) is an English rugby league footballer who plays as a forward for the Catalans Dragons in the Super League…Copied text…

Figure A4: **Example 1.**

Original Article

Isidore Mankofsky (born September 22, 1931, in New York City, New York) is an American cinematographer. He shot more than 200 educational movies for Encyclopaedia Britannica.

New Information

2021 Deaths in the United States
Isidore Mankofsky, cinematographer ("The Muppet Movie", "Somewhere in Time", "The Jazz Singer")

Deaths in March 2021
11 - Isidore Mankofsky, 89, American cinematographer ("The Muppet Movie", "Somewhere in Time", "The Jazz Singer")

The Parent Trap (franchise)
Additional crew and production details

Film	Crew / Detail
Parent Trap III	Joel McNeely, Isidore Mankofsky, Howard Kunin & Duane Hartzell

Ground Truth

Isidore Mankofsky (September 22, 1931 – March 11, 2021) was an American cinematographer, best known for his work on films such as "The Muppet Movie" (1979) and "The Jazz Singer" (1980)...Copied text…He died at his home in Los Angeles, California in March 2021 at the age of 89.

EDIT5

Isidore Mankofsky (September 22, 1931 – March 11, 2021) was an American cinematographer…Copied text…, and worked on "The Muppet Movie", "Somewhere in Time", and "The Jazz Singer".

Figure A5: **Example 2.**

Original Article

[0] "Shuggie Bain" is the debut novel by Scottish-American writer Douglas Stuart, published in 2020. [EDIT] [1] It tells the story of the youngest of the three children, Shuggie, growing up with his alcoholic mother, Agnes in the 1980s, in Thatcher-era Glasgow, Scotland. [EDIT] [2] (0) The novel won the 2020 Booker Prize. [EDIT] [3] (1) It was also a finalist for the 2020 National Book Award for Fiction.

New Information

James Kelman

Critical reception

In his essay "The Importance of Glasgow in My Work", he compares the presentation of working-class and Scottish characters with those of the traditional "upper-class" English protagonist: In 2020, Douglas Stuart on becoming the second Scottish writer to be awarded the Booker Prize, for his novel "Shuggie Bain", said that his life was changed by Kelman's win with "How Late It Was, How Late": "It is such a bold book, the prose and stream of consciousness is really inventive.

National Book Critics Circle Award

Finalists

2020 - John Leonard Prize

Kerri Arsenault, "Mill Town: Reckoning with What Remains" (St. Martin's), Karla Cornejo Villavicencio, "The Undocumented Americans" (One World), Raven Leilani, "Luster" (Farrar, Straus and Giroux), Megha Majumdar, "A Burning" (Knopf), Douglas Stuart, "Shuggie Bain" (Grove), Brandon Taylor, "Real Life" (Riverhead), C Pam Zhang, "How Much of These Hills Is Gold" (Riverhead)

Ground Truth

…Copied text… It tells the story of the youngest of the three children, Shuggie, growing up with his alcoholic mother, Agnes, in the 1980s, in a working-class Glasgow, Scotland. The novel was awarded the 2020 Booker Prize, making Stuart the second Scottish winner of the prize in its history, following James Kelman. "Shuggie Bain" was also a finalist for the 2020 National Book Award for Fiction and a finalist for the 2020 John Leonard Prize for Best First Book from the National Book Critics Circle.

EDIT5 - Controllable

Copied text... It tells the story of the youngest of three children, Shuggie, growing up with his alcoholic mother, Agnes, in [DELETED] thatcher-era Glasgow, Scotland. The novel won the 2020 Booker Prize, and was a finalist for the 2020 National Book Award for Fiction and the 2021 John Leonard Prize. It was also a finalist for the 2020 National Book Critics Circle Award.

Figure A6: **Using Control Codes.**

Instructions

Overview
The goal of this task is to collect evaluation data for a system that can automatically update Wikipedia articles from new information about the article's subject. The sections below provide the text of the *original passage* to be updated, a collection of *added information* about the article subject, and the text of the *updated passage*.

What we need from you
The issue we are faced with is that some of the udpated text may not be supported by the added information section. We need you to identify all of the unsupported information, and edit the article to remove unsupported text while preserving fluency. Please make sure that your edits only remove information; while you may need to write some text to ensure that the edited passage is fluent, no new facts should be added (even if they are supported).

The original passage and added information are below this box. We request that you first read the updated passage in the green box, and highlight any unsupported text. Then copy the contents from the green box to the orange box and edit them so that all of the text is supported.

If you have questions please do not hesitate to email: REDACTED

Original Passage - DO NOT CHANGE

```
[0] Joshua Christian Kojo King (born 15 January 1992) is a Norwegian professional footballer who plays
as a forward for Championship club Bournemouth and the Norway national team.
[1] King was signed by Manchester United from Vålerenga in 2008.
[2] After loan spells with Preston North End, Borussia Mönchengladbach, Hull City and Blackburn Rovers,
he signed permanently with Blackburn in January 2013, before switching to Bournemouth in May 2015.
[3] After representing Norway at under-15, under-16, under-18, under-19 and under-21 levels, King made
his senior international debut against Iceland in 2012, and scored his first international goal against
Cyprus later that year.
```

Added Information - DO NOT CHANGE

```
(0) 2020-21_AFC_Bournemouth_season Transfers - Transfers out - Table-0-29
| Date            | Position | Nationality | Name        | To      | Fee         | Ref. |
|-----------------|----------|-------------|-------------|---------|-------------|------|
| 2 February 2021 | SS       |             | Joshua King | Everton | Nominal fee |      |

(1) 2020-21_Everton_F.C._season Transfers - Transfers in - Table-0-6
| Date            | Position | Nationality | Name        | From        | Fee     | Team       | Ref. |
|-----------------|----------|-------------|-------------|-------------|---------|------------|------|
| 1 February 2021 | FW       |             | Joshua King | Bournemouth | Nominal | First team |      |

(2) Gullballen Winners - 2014-2017 - Table-0-3
| Year | Winner      | Club(s)     |
|------|-------------|-------------|
| 2017 | Joshua King | Bournemouth |

(3) 2020-21_Crawley_Town_F.C._season Review - January Nichols equalised from close range in the 59th
minute before Josh King scored Bournemouth's winner.

(4) 2020-21_Manchester_United_F.C._season Premier League McTominay restored the lead only for Dominic
Calvert-Lewin to equalise again in the final minute of stoppage time following Tuanzebe's foul on
Everton substitute and fellow United Academy graduate Joshua King.
```

Updated Passage - HIGHLIGHT UNSUPPORTED TEXT

```
(0)
(1) Joshua Christian Kojo King (born 15 January 1992) is a Norwegian professional footballer who plays
as a forward for Premier League club Everton and the Norway national team.
[1] King was signed by Manchester United from Vålerenga in 2008.
[2] After loan spells with Preston North End, Borussia Mönchengladbach, Hull City and Blackburn Rovers,
he signed permanently with Blackburn in January 2013, before switching to Bournemouth in May 2015.
(0)
(1) In February 2021, in a deadline day deal, he returned to the top flight with a move to Everton.
[3] After representing Norway at under-15, under-16, under-18, under-19 and under-21 levels, King made
his senior international debut against Iceland in 2012, and scored his first international goal against
Cyprus later that year.
```

Edited Passage - COPY FROM THE CELL ON THE LEFT AND EDIT

```
(0)
(1) Joshua Christian Kojo King (born 15 January 1992) is a Norwegian professional footballer who plays
as a forward for Premier League club Everton and the Norway national team.
[1] King was signed by Manchester United from Vålerenga in 2008.
[2] After loan spells with Preston North End, Borussia Mönchengladbach, Hull City and Blackburn Rovers,
he signed permanently with Blackburn in January 2013, before switching to Bournemouth in May 2015.
(0)
(1) In February 2021,he returned to Everton.
[3] After representing Norway at under-15, under-16, under-18, under-19 and under-21 levels, King made
his senior international debut against Iceland in 2012, and scored his first international goal against
Cyprus later that year.
```

Step 1

The section below above the text of the updated passage. Unchanged sentences from the original passage are in grey, while added or updated sentences are in black.

We have tried to automatically detect which pieces of added information justify the changed text. If a justification is detected, then the edited sentence will be prefaced with the delimiter of the added information.
For example:
(0) (1) Updated sentence
means that we think that added information 0 and 1 justify (at least some of the edit).

What to do for this column
- Highlight any text/evidence delimiters that are unsupported by the original passage or added information, using this red color (in the custom section).
- Do not edit the text.

Step 2

What to do for this column
- Copy the highlighted updated passage from the previous step.
- Edit text so that: a) any unsupported text is removed, and b) the passage is still fluent.
- Do not add any new information

Figure A7: Annotator Interface