NewsEdits 2.0: Learning the Intentions Behind Updating News

Anonymous ACL submission

Abstract

News articles are often published and republished. Their revision histories give us insights into the journalistic process and can assist in the development of computational journalism tools. They also make it challenging for large language models (LLMs) trained with news to reconcile conflicting, updating information. In this work, we release *NewsEdits 2.0*, based on Spangher et al. (2022)'s large corpus of news article revision histories. *NewsEdits 2.0* introduces a taxonomy of edit-intention categories, including coarse categories: Fact Updates, Stylistic Updates, Contextual/Narrative Changes and XX finer-grained categories. In the first part of our work, we collect ZZ human-labeled annotations on 600 revision-pairs, and show that we can model these categories using small, scalable ensemble models with high F1 score (YY). In the second part of our work we seek to model, given old versions of news articles: *will this article have fact updates? Will it have a style updates?* We show that, while pretrained LLMs fail at this task, fine-tuning can boost performance to YY accuracy. Finally, we show via a novel use-case, *Question Answering with outdated references*, that *NewsEdits 2.0* should play an important role for users.

1 Introduction

News is the "first rough draft of history" (Croly, 1943). It's information is both valuable and fluid, prone to changes, updates, and corrections. Spangher et al. (2022) gave insight into this fluidity by releasing *NewsEdits*, a large corpus of article revision histories. Authors asked: *"Which facts are uncertain and likely to be changed? Which events are likely to be updated? What voices and perspectives are needed to complete a narrative?"*

Intuitively, the sentence: *"Japan issued a tsunami advisory for the eastern coast"*, shown in Figure 1 is highly likely to update, while *"It hit the Fukushima nuclear plant"* is not, yet both

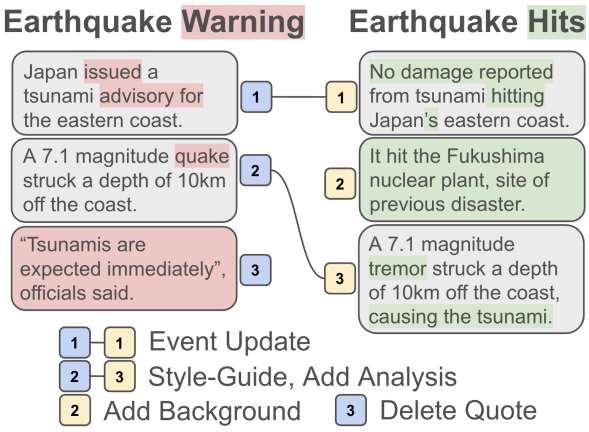

Figure 1: *NewsEdits 2.0*: We introduce a taxonomy of edit-types to characterize edits in revisions of news articles, an annotated dataset, models to predict intentions between versions, and a novel prediction task: *which facts are the most likely to update in a news article?* (shown here, fact-updates are "Event Update", "Quote Deletion" categories). We highlight that this prediction problem is particularly important, as LLMs otherwise confidently repeat outdated information.

sentences are equally likely in news corpora. As the role of large language models (LLMs) grows, and their use of news corpora for pretraining (Wu et al., 2021), summarization (Zhang et al., 2024) and prediction (Allaham and Diakopoulos, 2024) continues, the importance of understanding the fluidity of factual information in news articles increases. Spangher et al. (2022) showed that fine-tuned LLMs and expert journalists could predict whether an article would update and by how much. However, they left unaddressed the semantic quality of these updates: were they factual? Stylistic?

We address these shortcomings by improving upon *NewsEdits* in the following ways. First, we introduce a taxonomy of edit-intentions for journalistic edits working closely with 2 professional journalists and 1 professional copy-editor and building off work in adjacent domains (Yang et al., 2017; Zhang and Litman, 2015). Our schema, shown in

Figure 2: *NewsEdits 2.0*: Edit-Intentions Schema categories and their subcategories. In this work, we focus mainly on the *Factual Edit* category. See Appendix **??** for definitions for all categories.

Figure 2, has three categories: Factual Edits, Style Edits, Narrative/Contextual Edits. We enlist journalists to annotate 600 article revision pairs with XX total intentions. An ensemble approach, we find, reaches a prediction accuracy of ZZ F1 score.

We then frame a novel prediction problem: given *just* a news article, which sentences will have fact updates? Stylistic updates? Crucially, we find that leading pretrained LLMs do *not* answer either of these questions well. However, fine-tuning LLMs on silver-labeled data – created by applying our ensemble models to the revision histories of NewsEdits – enhances their performance in predicting Factual edits, achieving an F1-score of YY. *We do not observe a similar effect in predicting Style Edits.* Finally, we show that these predictions can have real impact. We highlight a use-case: Question Answering with outdated references. We simulate a case where an LLM using Retrieval Augmented Generation retrieves an outdated document. Without access to our predictions, the LLM answers confidently and wrongly XX% of the time.

In sum, our contributions are:

- We introduce *NewsEdits 2.0*, a large corpus of 1 million news articles and 4 million revision histories silver-labeled with edit intentions. We develop a schema with 3 coarse and 20 fine-grained categories, developed in conjunction with professional journalists, and train models to label these with ZZ F1.
- We introduce a challenging new task, *content evolution prediction*, where the goal is to predict which parts of a news article will update and how. Pretrained LLMs fail at this task, and while fine-tuning helps, performance still lags humans, indicating a challenging task.
- We show via a novel use-case, Question Answering with Outdated Documents, that a failure to address these shortcomings can result in decreased performance for leading LLMs.

The paper is structured as follows. In Section 2, we will introduce the schema (Section 2.1) and introduce our ensemble approach for revision histories (Section 2.4). Then, in Section **??**, we will discuss our edit-intention prediction problem (Section **??**), discuss our ability to model factual edits (Section **??**), and the implications for our use-case, Question Answering with outdated references (Section **??**). We also discuss our negative results around Style Edit predictions.

2 NewsEdits 2.0: Edit Intentions in Revision Histories

News articles update for different reasons, especially in a breaking news cycles where facts and events update quickly (Saltzis, 2012). We wish identify categories of edits that occur, in order to enable different investigations into these different update patterns. In other words, we describe the following update model:

$$p(l|D_i, D'_j, D, D') \quad (1)$$

Where l is a *reason* for updating (e.g. a "Correction" needs to be made, or an "Event Updated"), D and D' represent the older and newer versions of a news article, respectively, and D_i and D'_j are individual sentences where the update occurred. In prior work, these reasons are broadly described as "intentions", e.g. in Wikipedia (Rajagopal et al., 2022; Yang et al., 2017) and student-learner essays (Zhang and Litman, 2015). Although edit-intention schemas have been developed in these domains,

we suspect that the unique structural dynamics of newsroom publishing – where minute-by-minute updates are part of their business models (Rosenberg and Feldman, 2008) – result in markedly distinct editing patterns.

2.1 Edit Intentions Schema

We work with 2 professional journalists and 1 professional copy editor[1] to examine 50 revision-pairs sampled from *NewsEdits*, and iteratively expand and collapse our schema until it stays stable. Figure 2 shows our schema, which we organize into a hierarchy of coarse and fine-grained labels. While our coarse schema shares major overlaps with other revisions schema (e.g. Rajagopal et al. (2022) also identifies categories like "Fact Update" and "Style Update"[2]) we develop finer-grained categories based on intentions that are more important to journalists, like "Event Updates", "Quote Update" and "Sensitivity Consideration". In addition, we consciously incorporate existing theories of news semantics into our schema. For instance, our label "Event Updates" incorporates definitions of "events" from event detection (Doddington et al., 2004), while "Add Background" incorporates theories of news discourse (Van Dijk, 2013). "Add Quote" incorporates definitions from informational source detection (Spangher et al., 2023) and "Add Anecdote" incorporates definitions from argumentation (). Finally, the "Incorrect Link" category is an attempt to correct sentence pairs that were erroneously linked/unlinked by Spangher et al. (2022). See Section **??**. See Appendix **??** for a deeper discussion of the differences between our schemas.

2.2 Schema Annotation

We build an interface for annotators to provide fine-grained intention labels for news article sentence pairs. In the interface, annotators are shown definitions for each fine-grained edit category and the pair of news article revisions for context. For each sentence that was edited, annotators are asked to annotate the intention. To recruit annotators, we posted on two list-serves for journalism industry professionals[3]. We asked prospective applicants to describe their journalism experience, and selected this pool based on those having one or more year of professional editing experience. Then, we asked them to label revised sentences in 5 news articles, which we checked. We recruited 11 annotators who scored above 90% on these tests. See Appendix for more details about our annotators.

2.3 Technical Improvements over *NewsEdits*

Spangher et al. (2022) defined sentence edits, additions and deletions as the operations of change in revision histories and they report matching sentences across revision pairs with 89.5% F1. However, this error rate, we found, was noisy enough to warrant consistent negative feedback from our annotators. So, we examine *NewsEdits*'s sentence matches. A major category of error, we find, stems from poor sentence boundary detection (SBD). Poor SBD creates an abundance of sentence stubs and, because Spangher et al. (2022)'s sentence-matching method calculates a matching score normalized by the *shorter* sentence of a sentence pair[4], these stubs often over-match across revisions. To address this issue, we reprocessed the dataset from scratch. Instead of using SparkNLP for SBD[5], we use Spacy[6], which we qualitatively observe to be better. For word-matching, we use `albert-xxlarge-v2`[7]'s embeddings (Lan et al., 2019) instead of `TinyBert`(Jiao et al., 2019). These two steps, we find, increase our linking accuracy to 95%. We reprocess and release *NewsEdits* using our pipeline. Finally, we release a suite of visualization tools, based on D3[8], which are visually less cluttered than those released by Spangher et al. (2022), to enable further exploration of the corpus.

2.4 Modeling Edit Intentions

Edit intentions are labeled on the sentence-level, and each sentence addition, deletion or update is potentially multiply labeled. Furthermore, document-level context is important: for instance, understanding that Sentence 2 in Figure 1 ("*It hit the Fukushima nuclear plant, site of previous disas-*

[1] Collectively, these collaborators have over 50 years of experience in major newsrooms.

[2] Their schema actually contains "Wordsmithing", which is a close corollary to "Style Update".

[3] The Association of Copy Editors (ACES) https://aceseditors.org/ and National Institute for Computer-Assisted Reporting (NICAR) https://www.ire.org/hire-ire/data-analysis/.

[4] Authors design their *max-alignment* method in order to accommodate sentence splitting and merging edit-operations.

[5] https://sparknlp.org/api/com/johnsnowlabs/nlp/annotators/sbd/pragmatic/SentenceDetector.html

[6] https://spacy.io/, specifically, the `en_core_web_lg` model.

[7] https://huggingface.co/albert/albert-xxlarge-v2

[8] https://d3js.org/

Features	All		Fact		Style		Narrative	
	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro
Baseline, *fine-grained*	45.8	73.6	32.0	47.2	58.6	39.9	52.0	39.9
+ NLI	48.6	74.1	45.7	50.4	55.2	38.7	43.6	38.7
+ Event	46.7	74.1	39.0	49.0	59.3	41.4	41.7	41.4
+ Quote	46.3	72.8	49.8	54.7	31.9	28.0	42.4	28.0
+ Collapsed Quote	**51.2**	73.9	38.7	47.6	58.3	39.4	51.4	39.4
+ Discourse	45.8	75.1	37.7	49.6	63.8	44.6	43.2	44.6
+ Argumentation	48.9	73.6	37.1	47.9	57.1	37.7	53.5	37.7
+ Discourse & Event	46.3	74.3	38.9	49.9	62.1	42.2	42.4	42.2
+ Discourse & Argumentation	47.8	74.1	56.8	50.5	31.4	32.2	41.1	32.2
+ Argumentation & Event	50.0	75.1	38.0	48.6	46.4	44.9	58.5	44.9
+ Quote & Discourse	**51.2**	72.2	40.5	45.3	62.8	43.0	48.7	43.0
+ Collapsed Quote & Discourse	49.6	73.9	45.6	49.4	58.9	39.1	47.9	39.1
+ Collapsed Quote & NLI	45.4	72.8	41.9	50.4	46.7	31.2	39.3	31.2
+ Collapsed Quote & NLI & Event	49.0	73.8	44.9	48.9	57.4	37.0	44.0	37.0
+ All	47.2	73.6	40.0	49.7	58.6	36.0	43.5	36.0
Baseline, *coarse-grained*	49.4	56.7	46.6		65.1		10.4	
+ Discourse & Arg. (Best model, Fact)	65.4	70.7	59.4		66.2		49.2	

Table 1: Various F1 scores (%) on our test set of the fine-tuned LED model with different combinations of features. Fact/Style/Narrative F1 scores are computed on instances that contain the corresponding labels, whereas All F1 scores are derived from all instances.

ter.") is adding background is aided by the surrounding sentences contextualizing that a major tsunami event had just occurred.

Generative models have recently been shown to outperform classification-based models in document understanding tasks (Li et al., 2021; Huang et al., 2021). Inspired by this, we develop a sequence-to-sequence framework using the LongFormer-Encoder-Decoder (LED) architecture[9] (Beltagy et al., 2020) to predict the intent behind each edit. Specifically, our model processes the input $x = [D_i||D_j'||D||D']$. As shown in Figure 1, D_i or D_j' can optionally also be $\varnothing$, which corresponds to the other sentence being a deletion or an addition. The decoding target $y = [l_1||\ldots||l_n]$ is a concatenation of potentially multilabeled intention labels l_i for the edit from D_i to D_j'. The objective is maximum likelihood estimation, $\mathcal{L} = -\sum_i p(y_i|y_{<i}, x)$ where y_i denotes the i-th token in the concatenated label sequence. We include the context of the entire articles (i.e. D and D') after finding that additional context leads to a 4.91% improvement in Micro F1.

Experimental Variants and Results As discussed in Section 2.1, we developed our schema to bring together different theories of news semantics. We experiment with integrating labels from these other models published these domains. We use models from the following papers: *Discourse* (Spangher et al., 2021), *Quote-Type Labeling* (Spangher et al., 2023), *Event Detection* (Hsu et al., 2021), *Textual Entailment* (Nie et al., 2020) and *Argumentation* (). Labels generated from these external schema, denoted as f_{D_i} and $f_{D_j'}$, are appended to the model input $x = [D_i||D_j'||D||D'||f_{D_i}||f_{D_j'}]$. The efficacy of incorporating these features is reported in Table 1, where the Macro and Micro F1 scores increase by 5.46% and 1.46%, respectively. Model specifications such as input templates and additional schemas are illustrated in Appendix B.

2.5 Insights

We present a brief exploratory analysis, with more material shown in the appendix. We run the models trained in the last section over the entire corpus of 500,000 articles. First, we explore the correlation between syntactic edit categories (e.g. "addition", "edit", "delete") and the semantic categories that we have identified. As can be seen in Table 2, categories like Addition have far more Narrative and Factual updates than stylist updates; stylist updates, on the other hand, are far more likely to occur between sentences. This makes sense, stylistic updates are likely smaller, local updates, while Narrative and Factual updates might include more rewriting.

Next, we zero in on sentence-level updates, and we explore if there are types of content that are

[9] https://huggingface.co/allenai/led-base-16384

	Narrative	Fact	Style
addition	840329	358900	104
deletion	330039	21671	6088
edit	411292	102499	644243

Table 2: Counts of coarse-grained semantic edit types, broken out by syntactic categories (for fine-grained counts, see Appendix).

	Fact	Style	Other
Disaster	6.4	43.4	50.0
Elections	5.1	47.9	46.9
Environment	1.9	56.8	41.2
Labor	2.0	49.6	48.2
Other	3.7	50.7	45.5
Safety	4.7	46.6	48.6

Table 3: Distribution over update-types, across social-interest categories (Spangher et al., 2023).

more likely to have certain kinds of edits. We start by looking at high-level news categories, shown in Table 4. These are derived from training a classifier on CNN News Groups. "Politics" and "Sports" coverage are observed to have the highest level of fact-edits relative to other categories, while Style updates are prevalent in "Entertainment" pieces. Table 3 shows the kinds of edits in different categories of news determined "socially beneficial", by (Spangher et al., 2023) (we use a classifier released by the authors as well to derive these categories). This points the way to tools that might be helpful in these fields. Not surprisingly, even though "Fact" updates are rarer overall in sentence-level updates, they are more represented in Disaster and Safety categories.

We focus on Fact-updates in the next section, due to it's close relation to our demo use-case, and leave consideration of other edit categories to future work.

3 Content Evolution Prediction

3.1 Problem Statement

In the first half of the paper, we sought to describe intentions behind *observed* revision patterns, shown in Equation 1 and found that we could categories these patterns with reasonable performance. Now, we wish to leverage this groundwork to learn a predictive function:

$$p(l|D_i, D) \quad (2)$$

Where D_i and D are the *older* of a revision

	Fact	Style	Other
business	1.6	62.0	36.4
entertainment	3.3	65.5	31.1
health	2.1	61.0	36.9
news	2.8	57.0	40.2
politics	5.9	57.8	36.3
sport	3.5	59.3	37.2

Table 4: Distribution over update-types, across CNN section classifications.

history pair (or, in edge cases, the last version of a revision history sequence). In other words, we wish to describe how this article *might* change. This, we hypothesize, can allow us to take actions to help users as news unfolds (Section 4) and will help us learn patterns about the nature of the news event (whether it is fast-breaking or relatively stable) and role of the sentence in the story. Spangher et al. (2022) showed that, to some degree, structural predictions could be made about how a news article developed across time. They modeled whether an article *would* update or not with F1 > .77. And they showed that expert journalists were surprising good at predicting how *much* and *where* an article would be update. However, authors stopped at this "syntactic" analysis. In this work, we go a step further: with the semantic understanding of edits introduced in the prior section, we try to predict *how* information will change.

3.2 Dataset Construction

Because Spangher et al. (2022) already demonstrated "syntactic" predictability, discussed in the prior section, we can safely narrow our focus to articles that we know have substantial updates. We sample a set of 500,000 articles from *NewsEdits* that have > 10% sentences added and > 5% sentences deleted. Then, we use models developed in Section 2.4 to produce silver-standard labels. In other words, we assign labels l using both versions of a revision pair (Equation 1) and then we discard D', D'_j and try to predict l using *just* D, D_i (Equation 2).

In order to prevent label leakage, we perform a chronological split of our dataset, splitting the earliest 80% of articles for training and the next 10% as the development set, and the most recent 10% as the test set. To keep computational and cost requirements reasonable and reproducible, we sample 16,000 sentences for the training set and 2,000 each for the development and the test set. In early experiments, we noticed that many fine-grained labels were too infrequent to model well, so we

Model	Features	Fact F1	None F1	Macro F1	Weighted F1
GPT-3.5	S	11.3	79.1	30.4	74.2
	DC	3.4	91.8	32.2	85.2
	FA	7.9	91.1	49.8	85.4
GPT-4	S	11.1	66.3	38.9	62.4
	DC	14.8	88.8	52.7	84.1
	FA	15.4	90.6	53.2	84.9
FT Longformer	S	21.2	92.3	57.4	87.0
	DC	22.3	93.0	87.8	87.4
	FA	25.4	91.4	58.0	86.4
Human Performance	S	41.2	75.3	58.6	69.2

Table 5: Individual F1 scores and macro and weighted F1 scores (%) on the golden test set for various evaluated models. S: sentence-only, DC: direct context, FA: full article.

switched to predicting coarse-grained labels. As shown in Section 2.5, this classification problem is highly imbalanced: there are many more sentences that are not updated and of those that are, Style and Background/Narrative categories are more common. Thus, we balance the training dataset to have an equal number of classes for training. We sample from the true distribution for the development and test set. This yields a test set with 1,654 nones; 211 fact-updates; and 135 style-updates.

3.3 Experiments

We take two different approaches: (1) **Article context** We hypothesize that the broader article context is necessary to predict sentence-level update semantics, as sentences play a discursive role in the larger story (Van Dijk, 2013). Thus, we design an experiment to predict with (i) only the target sentence, (ii) with the direct context (one sentence before and after), and (iii) with the full article. We evaluate zero-shot approaches (prompted `gpt-3.5-turbo` and `gpt-4` and longformer models and `gpt-3.5-turbo` models finetuned each of the configuration above with the training set. The longformer is trained with the same approach as the silver-label prediction step from Section 2.4 and `gpt-3.5-turbo` is trained using the OpenAI API. (2) **Topic and Dataset Restriction** Next, we tried to model different subsets of our dataset. As shown in Section 2.5, various types of content have different patterns of factual or stylistic update patterns. However, we find negative results: attempting to train models with a training and test set that only contain specific topics, like "Disaster" or "Safety", does not show any performance improvements. For both approaches, we evaluate their performance on the same set of documents D_{test}^{gold}, which were part of the test set of our annotation, described in Section 2.2.

3.4 Results

Results are shown in 5. As can be seen, performance is overall for detecting fact-updates. However, we do observe performance increases from training the longformer model, so to some degree this task is learnable. We recruit a former journalist with years of experience in newsrooms to provide human evaluations as an upper bound. With some observation of the training data, the journalist is able to determine that certain kinds of information: e.g. death counts and other statistics, present-tense or future tense events, etc. are highly likely to change. At 41.2 F1-score, the journalist sets an upper bound, but not a very high upper bound.

We next hypothesize that the middle of the distribution is actually very noisy: many things may or may not have Fact updates even if they look very similar, because of case-by-case journalistic decision-making. So, we explore performance in the high-precision region: the region of the probability distribution where, we assume, edits are so necessary that noise is reduced. Figure 3 shows this exploration. As we restrict the pool of documents, we increase the performance.

	Easy			Medium			Hard		
	W. F1	Macro F1	Avg.	W. F1	Macro F1	Avg	W. F1	Macro. F1	Avg.
Baseline #1	55.9	35.8	55.9	8.8	8.1	8.8	38.8	28.0	38.8
Baseline #2	52.9	**49.6**	52.9	90.0	47.4	90.0	64.7	54.0	64.7
Experiment	**59.4**	48.9	**59.4**	**90.6**	61.1	**90.6**	**67.1**	**62.4**	**67.1**
Oracle	57.6	47.7	57.6	90.0	**63.3**	90.0	66.5	61.1	66.5

Table 6: A use-case we apply NewsEdits2.0 to: predicting when to abstain from factual question-answering, based on updated material. We generate questions in different categories (easy, medium, hard

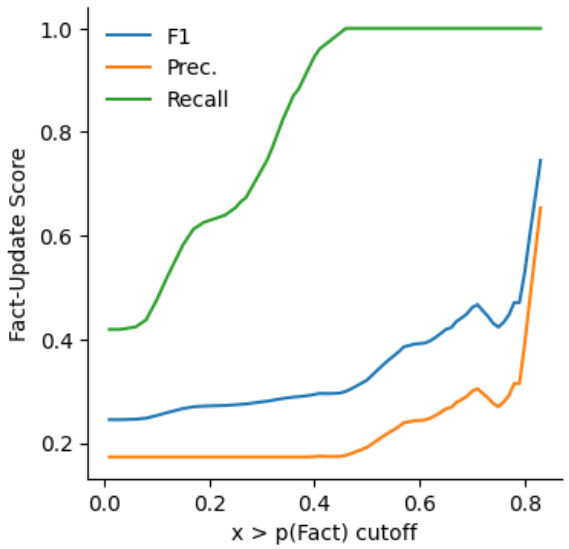

Figure 3: Performance of Fact-update model increases as we increasingly focus on a pool of documents that are categorized as high-likelihood under the model. In otherword, the model truly shines in the high-precision, high-probability realm.

4 Use-Case: Question Answering with Outdated Documents

We consider a useful application of the Fact edits. Kasai et al. (2022) raised an interesting observation: "*GPT-3 tends to return outdated answers when retrieved documents [are outdated]. Can an open domain QA system identify such unanswerable cases?*" We explore whether our predictions for Fact label, generated in the previous section from modeling Equation 2, can be used downstream to prompt GPT to look out for outdated information.

We set up the following experiment. We want to see whether we can induce behavior in GPT where GPT *abstains* from answering a question if it suspects that the information it's basing it's answer on might be out-of-date. We prompt GPT with (1) an *outdated* sentence from an old version of a news article and (1) a question that relies on information in the sentence. *If* a newer version of the article exists and *if* the information in the newer version would lead to a different answer, then we consider that providing a confident answer to the question is a failure. Consider the following example:

> **Old sentence**: "The White House is on lockdown after a passenger vehicle struck a security barrier."
> **New sentence**: 'The White House was on lockdown for about an hour Friday after a passenger vehicle struck a security barrier.'
> **Question**: "If I visit the White House right now, will I get turned away?"

Remember: GPT only has access to the old sentence, so it is likely expecting to be able to answer the question. However, confidently answering "Yes, you will be turned away" is considered *wrong* in this scenario, because a newer version of the article exists which suggests that lockdown is already over. Our models, on the other hand, are well-primed to detect that this sentence contains a short-duration temporal event which is stated in the present tense, so it is highly that this sentence will update (see Table 9 for more examples of high-probability sentences). So, we design the following trials. We take pairs of sentences in the gold test set of our annotated data where an update occurred, and we ask GPT4 to generate 15 questions per pair of sentences. 5 of these are "Easy" questions: they should be directly answerable from the old version, and not necessarily in conflict with the new version. Another 5 are "Medium", they are questions that might be in conflict with a potential newer version. Finally, the last 5 are "Hard", they are *definitely* in conflict with a newer version, because GPT4 is shown both versions and explicitly told to ask questions that fit this criteria (for all prompts, see Appendix D). We test the following baselines:

	Easy	Medium	Hard
Baseline #1	0.0	0.0	0.0
Baseline #2	30.0	98.8	87.1
Experiment	10.6	95.9	74.1
Oracle	12.4	94.1	75.9

Table 7: Likelihood of refraining. In general, we wish to refrain only when we need to. Over-refraining is bad.

Baseline #1: Vanilla We feed in a vanilla, basic prompt to GPT3.5, without alerting it to any possibly outdated material.

Baseline #2: Uniform We formulate a prompt that warns GPT3.5 that some information might be outdated, and to refrain from responding if it things it is. However, this prompt is the same for all questions, so GPT has to rely on it's own internal check to detect outdated information.

Experiment : Here, we feed in probabilities from out prediction model, binned into "low", "medium", "high" risk of being outdated. In other words, we tell GPT in the prompt that we suspect there is a high likelihood for the sentence being outdated, and to refrain from answering if the question directly potentials to information that might fall into that category.

Oracle : We feed in labels (in this case, gold labels) about whether a fact-updated *will* occur in the next version of the article. We keep the phrasing the same as in the experimental version. This is designed to give us an upper bound.

We evaluate performance of each prompting strategy as follows: we feed GPT4 the sentence pairs and the questions that were generated, and we ask: Is this question answerable given JUST the old sentence? Is the answer, using the old sentence, factually consistent with the information presented in the revised sentence? If the answer is yes to both, then GPT should answer confidently. If one of the answers is no, then we want GPT to refrain from answering. We count the number of times GPT refrains. Every time it refrains when it *should* be refraining is a success, otherwise is a failure.

Our results are shown in Table 6 and 7. Interestingly, and perhaps unexpectedly, the experimental variant does as well if not better than the oracle. Perhaps the granularity of the prediction score helps GPT make a better assessment of the likelihood of update; perhaps our gold labels are a bit overly broad. As expected, **Baseline #2** has a strong performance (Table 6), but at the cost of far more refrains, show in Table 7.

5 Related Work

A significant contribution of this work, we feel, is the semantic tools to make better use of existing high-quality datasets, and to make revisions history in journalism more accessible. Two works that analyze news edits to predict article quality (Tamori et al., 2017; Hitomi et al., 2017) do not release their datasets. In previous corpora, the nature of edits are primarily argumentative or corrective. However, news articles very often cover updating events. This difference has important implications for the kinds of edits we expect in our corpora.

Many tasks have benefited from studying **Wikipedia Revisions**, like text simplification, textual entailment (Dagan et al., 2005), discourse learning (Van Dijk, 2013) and grammatical error correction (**?**). However, most tasks focus on word-level edit operations to explore sentence-level changes. Research in **Student Learner Essays** focuses on editing revisions made during essay-writing (Zhang and Litman, 2015). Researchers categorize the intention and effects of each edit, but do not try to predict edits.

6 Conclusion

We introduce in this work NewsEdits2.0: a deeper semantic understanding of the editing decisions that journalists make. We have introduced a novel schema, grounded in theory in conjunction with professional journalists. We operationalized this schema and modeled it, showing that in combination with other advances in computational journalism, we can achieve higher performance. Then, we applied this models to create data to train predictive models. We extensively explore Fact-Updates, towards our use-case: prompting GPT to be aware of outdated information. We found that we were able to model the high-precision region of fact-updates well. We were able to prompt GPT with outputs from our model and achieve a suitable balance between refraining from commenting and serving users. We look forward in future work looking at the different edit categories: Style Edits and Background Edits. Our work here lays a firm groundwork for these directions.

7 Ethical Considerations

7.1 Dataset

NewsEdits is a publicly and licensed dataset under an `AGPL-3.0 License`[10], which is a strong "Copy-Left" license.

Our use is within the bounds of intended use given in writing by the original dataset creators, and is within the scope of their licensing.

7.2 Privacy

We believe that there are no adverse privacy implications in this dataset. The dataset comprises news articles that were already published in the public domain with the expectation of widespread distribution. We did not engage in any concerted effort to assess whether information within the dataset was libelious, slanderous or otherwise unprotected speech. We instructed annotators to be aware that this was a possibility and to report to us if they saw anything, but we did not receive any reports. We discuss this more below.

7.3 Limitations and Risks

The primary theoretical limitation in our work is that we did not include a robust non-Western language source. As our work builds off of NewsEdits as a primary corpora, it contains English and French.

This work should be viewed with that important caveat. We cannot assume *a priori* that all cultures necessarily follow this approach to breaking news and indeed all of the theoretical works that we cite in justifying our directions also focus on English-language newspapers. One possible risk is that some of the information contained in earlier versions of news articles was updated or removed for the express purpose that it was potentially unprotected speech: libel, slander, etc. Instances of First Amendment lawsuits where the plaintiff was successful in challenging content are rare in the U.S. We are not as familiar with the guidelines of protected speech in other countries.

We echo the risk of the original *NewsEdits* authors: another risk we see is the misuse of this work on edits for the purpose of disparaging and denigrating media outlets. Many of these news tracker websites have been used for good purposes (e.g. holding newspapers accountable for when they make stylistic edits or try to update without giving notice). But we live in a political environment that is often hostile to the core democracy-preserving role of the media. We focus on fact-based updates and hope that this resource is not used to unnecessarily find fault with media outlets.

[10] https://opensource.org/licenses/AGPL-3.0

7.4 Computational Resources

The experiments in our paper require computational resources. Our models run on a single 30GB NVIDIA V100 GPU or on one A40 GPU, along with storage and CPU capabilities provided by our campus. While our experiments do not need to leverage model or data parallelism, we still recognize that not all researchers have access to this resource level.

We use Huggingface models for our predictive tasks, and we will release the code of all the custom architectures that we construct. Our models do not exceed 300 million parameters.

7.5 Annotators

We recruited annotators from professional journalism networks like the NICAR listserve, which we mention in the main body of the paper. All the annotators consented to annotate as part of the experiment, and were paid $1 per task, above the highest minimum wage in the U.S. Of our 11 annotators, all were based in large U.S. cities. 8 annotators identify as white, 1 as Asian, 1 as Latinx and 1 as black. 8 annotators identify as male and 3 identifies as female. This data collection process is covered under a university IRB. We do not publish personal details about the annotations, and their interviews were given with consent and full awareness that they would be published in full.

7.6 References

References

Mowafak Allaham and Nicholas Diakopoulos. 2024. Supporting anticipatory governance using llms: Evaluating and aligning large language models with the news media to anticipate the negative impacts of ai. *arXiv preprint arXiv:2401.18028*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

H.D. Croly. 1943. *The New Republic*. v. 108. Republic Publishing Company.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Yuta Hitomi, Hideaki Tamori, Naoaki Okazaki, and Kentaro Inui. 2017. Proofread sentence generation as multi-task learning with editing operation prediction. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 436–441.

I Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, et al. 2021. Degree: A data-efficient generation-based event extraction model. *arXiv preprint arXiv:2108.12724*.

Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. 2021. Document-level entity-based extraction as template generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5257–5269, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime qa: What's the answer right now? *arXiv preprint arXiv:2207.13332*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Dheeraj Rajagopal, Xuchao Zhang, Michael Gamon, Sujay Kumar Jauhar, Diyi Yang, and Eduard Hovy. 2022. One document, many revisions: A dataset for classification and description of edit intents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5517–5524.

Howard Rosenberg and Charles S Feldman. 2008. *No time to think: The menace of media speed and the 24-hour news cycle*. A&C Black.

Kostas Saltzis. 2012. Breaking news online: How news stories are updated and maintained around-the-clock. *Journalism practice*, 6(5-6):702–710.

Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 498–517.

Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2023. Identifying informational sources in news articles. *arXiv preprint arXiv:2305.14904*.

Alexander Spangher, Xiang Ren, Jonathan May, and Nanyun Peng. 2022. Newsedits: A news article revision dataset and a novel document-level reasoning challenge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–157.

Hideaki Tamori, Yuta Hitomi, Naoaki Okazaki, and Kentaro Inui. 2017. Analyzing the revision logs of a Japanese newspaper for article quality assessment. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 46–50, Copenhagen, Denmark. Association for Computational Linguistics.

Teun A Van Dijk. 2013. *News as discourse*. Routledge.

Chuhan Wu, Fangzhao Wu, Yang Yu, Tao Qi, Yongfeng Huang, and Qi Liu. 2021. Newsbert: Distilling pre-trained language model for intelligent news application. *arXiv preprint arXiv:2102.04887*.

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.

Fan Zhang and Diane Litman. 2015. Annotation and classification of argumentative writing revisions. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pages 133–143.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

	Addition	Deletion	Edit
Add/Delete/Update Background	806909	329652	411025
Add/Delete/Update Quote	303451	17995	46300
Incorrect Link	191022	125362	237437
Other (Please Specify)	84646	66929	65077
Add/Delete/Update Event Reference	37409	3645	56098
Add/Delete/Update Analysis	33426	390	268
Add/Delete/Update Eye-witness account	9772	0	3
Add/Delete/Update Source-Document	6639	2	28
Add/Delete/Update Information (Other)	1058	13	3
Additional Sourcing	573	15	29
Tonal Edits	102	6000	616514
Emphasize/De-emphasize Importance	1	32	1076
Syntax Correction	1	2	21729
Emphasize/De-emphasize a Point	0	53	1668
Simplification	0	0	3
Style-Guide Edits	0	1	3253
Correction	0	1	47

Table 8: Counts of fine-grained semantic edit types, broken out by syntactic categories

A Appendix

B Details of the LED Model

In this section, we describe the specifications of the LED model described in Section 2.4.

B.1 Input Template

The input to the LED model is shown below:

```
Predict the edit intention from
version 1 to version 2.
Version 1: SOURCE_SENTENCE
Version 2: TARGET_SENTENCE
Version 1 Document: SOURCE_DOCUMENT
Version 2 Document: TARGET_DOCUMENT
```

Here, **SOURCE_DOCUMENT** (D) and **TARGET_DOCUMENT** (D') refer to the newer and older articles, while **SOURCE_SENTENCE** (D_i) and **TARGET_SENTENCE** (D'_j) represent a sentence with these articles.

B.2 Additional Schema

NLI We use textual entailment from (Dagan et al., 2005), which consists of *Entail*, *Contradict* and *Neutral*. These categories indicate whether two pieces of information refute each other, complement each other, or are neutral.

Event Detection

Argumentation

Quote

C Annotation Details

In this section, we provide details of the annotation process, such as annotation guidelines and task allocation.

C.1 Annotation Guidelines

To complete the task, look at each sentence: if it's been added, updated, or deleted between drafts, try to determine based on your knowledge of the journalistic editing process why this was done.

You can specify multiple intentions for each add/delete/edit operation. Please also pay attention to when sentences are moved around in a document (i.e. if that was done to emphasize or de-emphasize that sentence), and when there might be errors to how we are linking sentences.

We devised these in consultation with professional journalists. However, if you are consistently annotating edits with "Other" (i.e. we are missing something in our schema), please let us know!

Fact Edits:

- **Delete/Add/Update Eye-witness Account**: The writer deletes/adds/updates the contents for the events being described. This can either take the form of a quote (in which case this edit should be paired with a Quote Update), or a first-person account by the journalist.

- **Delete/Add/Update Event**: There is a change to some event in the world that the article covers and the article needs to be updated to reflect this. Usually, there are changes to the verbs in the article, but this can also include increased death counts, stock-market changes, etc.

- **Delete/Add/Update Source-Doc**: Additional written documents have been released by a government or company that warrant deletion/inclusion/update of the content of the article. For example, additional information included in an SEC filing, quarterly earnings report, IPCC report, etc.

- **Correction**: There are factual errors in the original version. The new version corrects the error.

- **Delete/Add/Update Quote**: There is an addition, editing or deletion of quotes in the article. Or, a quote from one person is swapped for a quote from another. Sometimes these updates are made with other intentions (e.g. to include a punchier quote, in which case it would also be a Preferential Edit. In these cases, please use the "+" button to add another intention dropdown.)

- **Additional Sourcing (Other)**: The new version includes evidence of new sources for additional information, usually added for confirmation purposes. Note that this is different from Quote Update or Document Update since Additional Sourcing doesn't have to result in a new quote or document reference. Can simply be an indication that the journalist obtained new evidence.

- **Additional Information (Other)**: This edit intention is applied when the new version of the article includes details or context not present in the original version, which doesn't necessarily fall under specific updates like eyewitness accounts, event changes, document updates, or sourcing alterations.

Style Edits:

- **Simplification**: educes the complexity or breadth of discussion. This edit might also remove information from the article.

- **Emphasize/De-emphasize Importance**: The sentence is moved up or down in the document in order to make the sentence MORE/LESS prominent, or to emphasize/de-emphasize it's connection to the events being described in another sentence.

- **Define term**: The author provides meaning or differentiation to a term or concept that might be unknown to the reader. Note that this intention is DIFFERENT from the Background intention, which is more about providing context, e.g. historical or geographic context for a person, company, or place.

- **Style-Guide Adherence**: Edits that are made specifically to address a formal style guide (when in doubt, defer to the Associated Press style-guide). The first version violates the style guide and the revised version fixes it.

- **Syntax Correction**: Improve grammar, spelling, or punctuation. These are strictly to correct errors in syntax, not **Preferential Edits**. And, they need not be adhering to a formal style-guide (when a **Syntax Correction** is also adhering to a **Style Guide**, please use the "+" button to add another intention dropdown and annotate both).

- **Tonal Edits**: The journalist or copy-editor made the edits due to a specific personal or artistic preference. Use your intuition here: these are usually edits that introduce punch, elegance or scenery. These edits often also have the effect of some other edit intention, see the example, but cannot be fully ascribed to other aims.

- **Sensitivity Consideration**: The journalist rewrote the sentence because the original version is inappropriate/ may be considered insensitive.

Narrative Edits:

- **Delete/Add/Update Analysis**: The writer deletes/adds/updates inferences from the presented information. These can be in the form of analyses, expectations, or deeper understandings. These are usually forward-looking rather than Background information, which is usually past-looking.

- **Delete/Add/Update Background**: Delete/add/update contextualizing information to the article to help readers understand the history, geography or significance of a term, personal, place or company. Note that contextualizing information is not analysis, expectations, or projections, which would fall into the Analysis intention category.

- **Delete/Add/Update Anecdote**: The writer deletes, adds, or updates a brief, revealing account of a person or event. This can be a personal story, a particular incident, or a narrative snippet that exemplifies a point or adds a humanizing or illustrative dimension to the news piece. These anecdotes may serve to engage the reader's interest, illuminate a fact, or provide a real-world example of abstract concepts.

Others:

- **Incorrect Link**: This refers to an error in our original linking of sentences. We have linked two sentences that should NOT be linked. This only pertains to 'Edit'ed or 'Unchanged' sentences. Sentences should not be linked if they are entirely unrelated — they have substantially different syntax, intent, and purpose — and, by error, our algorithm said they were. If you identify an **Incorrect Link** AND there are more than one links, please specify (A) the index of the sentence in the other version that it should NOT be linked to via the dropdown (B) any other intention ascribed to this pair (i.e. Fact Deletion).

C.2 Annotation Interface

Figure 4 shows the annotation interface for our task.

C.3 Annotation Task Distribution

In Figure 5, we show the portion of annotation tasks assigned to each worker.

D Prompts for Use-Case

D.1 Question-Asking Prompts

Easy I will give you a sentence and you will give me an answer. It should be timely and related to the facts in the sentence. It should be a question that could go stale, especially for ongoing events, or facts like death counts that might update.

Here are some examples: example 1: sentence: "WASHINGTON (AP) – The White House is on lockdown after a passenger vehicle struck a security barrier." question: "Is the White House currently in lockdown – if I visit, will I get turned away?"

example 2: sentence: "The death count from the street bombing is 49 injured, 2 killed so far." question: "How many people have died so far?"

example 3: sentence: "The construction work left the bridge badly damaged and unsafe for passengers and is expected to remain so for days." question: "What route should I take? The bridge is the quickest way to work."

Ok, now it's your turn. Ask 5 different questions, output in a list. Don't say anything else. sentence:

Easy I will give you a sentence and you will give me 5 different questions. It should be directly answerable by the sentence.

Here are some examples: example 1: sentence: "WASHINGTON (AP) – The White House is on lockdown after a passenger vehicle struck a security barrier." question: "What did the vehicle strike?"

example 2: sentence: "The death count from the 42nd street bombing is 49 injured, 2 killed so far." question: "Where did the bombing take place?"

example 3: sentence: "The construction work left the bridge badly damaged and unsafe for passengers and is expected to remain so for days." question: "What kind of work was being done?"

Ok, now it's your turn. Ask 5 different questions, output in a list. Don't say anything else. sentence:

Hard I will give you two sentences from an updating news article and you will give me 5 different questions. They should ideally focus on information that changes in between the sentences. So, if someone were to just look at the old sentence and you asked them your question, they would get it wrong.

Here are some examples: example 1: old sentence: "WASHINGTON (AP) – The White House is on lockdown after a passenger vehicle struck a security barrier." new sentence: 'WASHINGTON (AP) – The White House was on lockdown for about an hour Friday after a passenger vehicle struck a security barrier.' question: "Is the White House currently in lockdown – if I visit, will I get turned away?"

example 2: old sentence: "ISTANBUL (AP) – An earthquake with a preliminary magnitude of 6.2 shook western Turkey and the Greek island

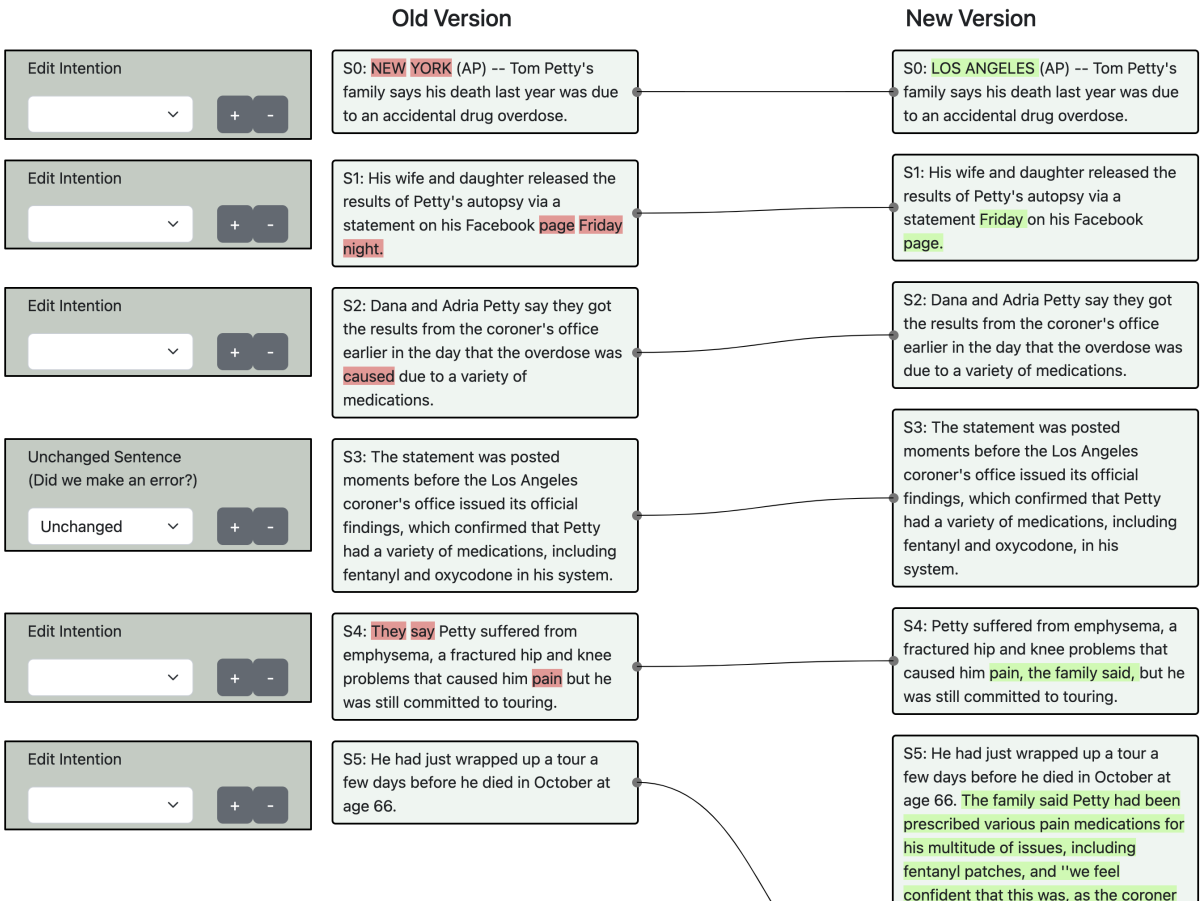

Figure 4: The interface for annotating edit intentions.

of Lesbos Monday, scaring residents and damaging buildings." new sentence: "ISTANBUL (AP) – An earthquake with a preliminary magnitude of 6.2 shook western Turkey and the Greek island of Lesbos on Monday, injuring at least 10 people and damaging buildings, authorities said." question: "Was anyone injured?"

example 3: old sentence: "Turkey's emergency management agency said there were no reports of casualties in the country." new sentence: "Turkey's emergency management agency said there were no reports of casualties and has dispatched emergency and health teams, and 240 family tents to the area as a precaution." question: "Is the Turkish emergency management doing anything as a precaution?"

Ok, now it's your turn. Ask 5 different questions, output in a list. Don't say anything else. old sentence: $\{\texttt{old}_s entence\} new sentence$:

Experimental Prompt `You are a helpful assistant who answers questions based on this news information: orig_sentence`

`We give this a {outdated_threshold` $\in \{high, medium, low\}$ `} chance of there being a fact update in this sentence. That might mean some new information, updating information. Answer cautiously and do not give the user wrong/outdated information. If the user's question looks like it will still be relevant even if the facts change, answer it directly. If the user's question looks like it will be outdated, say "I don't have the most up-to-date information" and that's it. Say nothing else. Do NOT say "I don't have the most up-to-date information" AND something else. Keep our estimate in mind.`

Baseline 1 `You are a helpful assistant who answers questions based on this news information: {orig_sentence}`

`Try to directly answer the users question and say nothing else.`

Baseline 2 `You are a helpful assistant who answers questions based on this news information: {orig_sentence}`

`This sentence might go out of date.`

Answer cautiously and do not give the user wrong/outdated information. If the user's question looks like it will still be relevant even if the facts change, answer it directly. If the user's question looks like it will be outdated, say "I don't have the most up-to-date information" and that's it. Say nothing else. Do NOT say "I don't have the most up-to-date information" AND something else.

Oracle You are a helpful assistant who answers questions based on this news information: {orig_sentence}

This sentence {oracle} have a major fact update. That might mean some new information, updating information. Answer cautiously and do not give the user wrong/outdated information. If the user's question looks like it will still be relevant even if the facts change, answer it directly. If the user's question looks like it will be outdated, say "I don't have the most up-to-date information" and that's it. Say nothing else. Do NOT say "I don't have the most up-to-date information" AND something else.

D.2 Evaluation Prompts

You are a helpful assistant. You will be shown an old sentence, a revised sentence, and a user-question. you will answer the following 2 questions: 1. Is this question answerable given JUST the old sentence? Answer with "yes" or "no". Do not answer anything else. If the answer to 1 was yes, then proceed to the second question, otherwise respond to question 2 with n/a 2. Does the question ask about something that is factually consistent with the information presented in the revised sentence? Answer with "yes", "no" or "n/a." Do not answer with anything else.

E Additional EDA

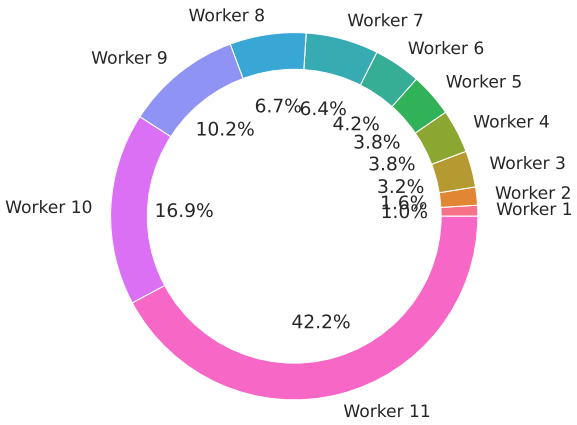

Figure 5: The portion of annotation tasks assigned to each worker.

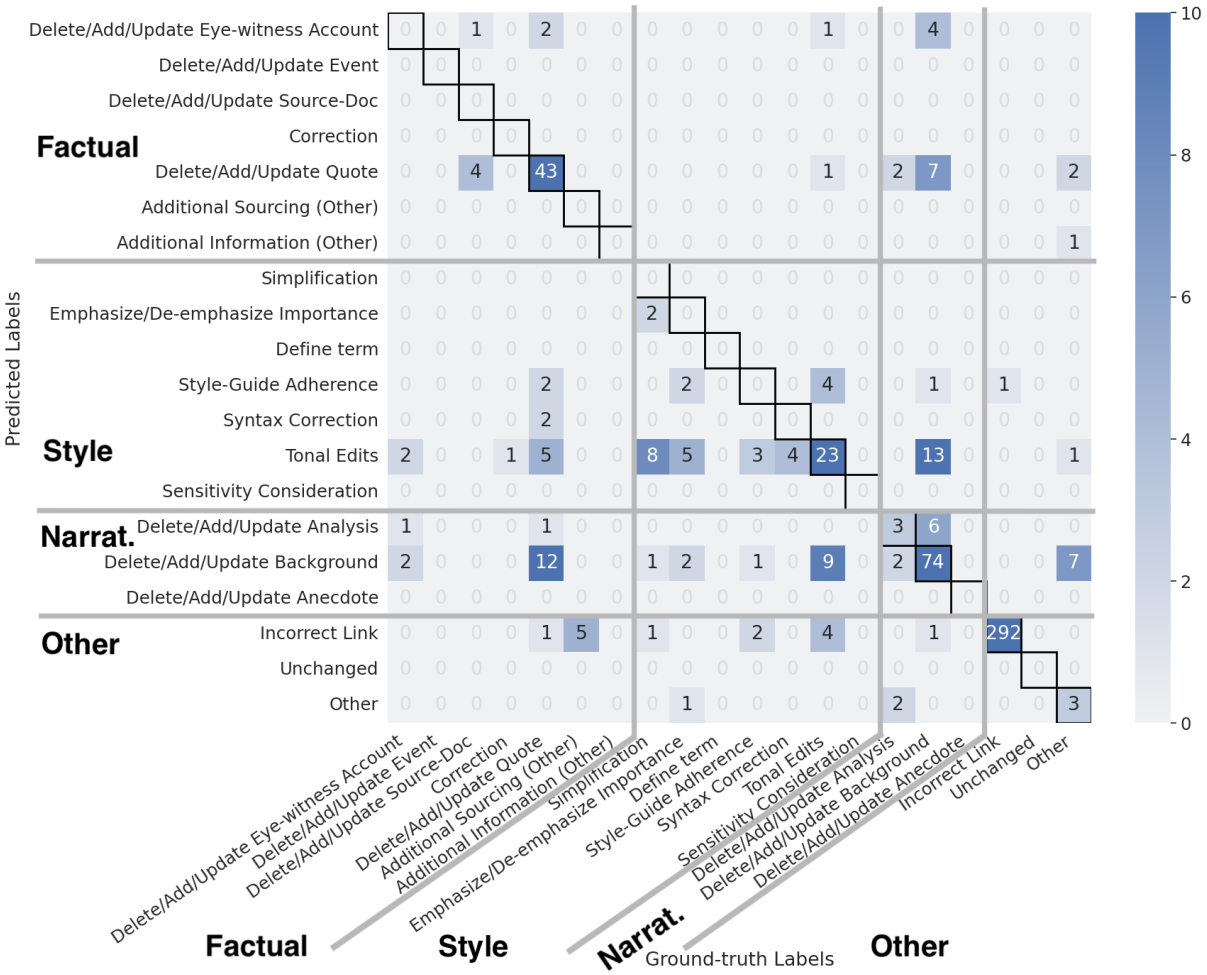

Figure 6: The confusion matrix for the LED model trained with Discourse and Argumentation features.

Top Predictions for Content Evolution Prediction, $p(l$ = Fact Update$\vert D_i, D)$
The company takes this recommendation extremely seriously," it said in a statement.
KABUL, Afghanistan — An Afghan official says a powerful suicide bombing has targeted a U.S. military convoy near the main American Bagram Air Base north of the capital Kabul.
WASHINGTON — The U.S. carried out military strikes in Iraq and Syria targeting a militia blamed for an attack that killed an American contractor, a Defense Department spokesman said Sunday.
Mr. Causey, who reported his concern to authorities, was not charged in the indictment, which a grand jury returned last month, and did not immediately comment.
His trial has not yet started.
MEXICO CITY — A fiery freeway accident involving a bus and a tractor-trailer killed 21 people in the Mexican state of Veracruz on Wednesday, according to the authorities and local news outlets.
The indictment accuses Mr. Hayes, a former congressman, of helping to route $250,000 in bribes to the re-election campaign of Mike Causey, the insurance commissioner.
No Kenyans died in the attack, Kenya's military spokesman Paul Njuguna said Monday.
Mr. Manafort, 70, will most likely be arraigned on the new charges in State Supreme Court in Manhattan later this month and held at Rikers, though his lawyers could seek to have him held at a federal jail in New York, the people with knowledge said.
Officials said attackers fired as many as 30 rockets in Friday's assault.
KABUL, Afghanistan — Gunmen attacked a remembrance ceremony for a minority Shiite leader in Afghanistan's capital on Friday, wounding at least 18 people, officials said.
BEIRUT — A senior Turkish official says Turkey has captured the older sister of the slain leader of the Islamic State group in northwestern Syria, calling the arrest an intelligence "gold mine. "
Paul J. Manafort, President Trump's former campaign chairman who is serving a federal prison sentence, is expected to be transferred as early as this week to the Rikers Island jail complex in New York City, where he will most likely be held in solitary confinement while facing state fraud charges, people with knowledge of the matter said.
The watchdog, the Securities and Exchange Surveillance Commission, said Tuesday it made the recommendation to the government's Financial Services Agency on the disclosure documents from 2014 through 2017.
There are no immediate reports of casualties.
It said the U.S. hit three of the militia's sites in Iraq and two in Syria, including weapon caches and the militia's command and control bases.
The rebel group did not immediately comment.
Kep provincial authorities later announced a total of five dead and 18 injured.
QUETTA, Pakistan — Attackers used a remotely-controlled bomb and assault rifles to ambush a convoy of Pakistani troops assigned to protect an oil and gas facility in the country's restive southwest, killing six soldiers and wounding four, officials said Tuesday.
WASHINGTON — Senator Bernie Sanders of Vermont raised $18.2 million over the first six weeks of his presidential bid, his campaign announced Tuesday, a display of financial strength that cements his status as one of the top fund-raisers in the sprawling Democratic field.

Table 9: Sample of the most likely fact-update sentences, as judged by our top-performing model. Top predictions reflect a combination of statistics, recent or upcoming events, and waiting for quotes.

Lowest Predictions for Content Evolution Prediction, $p(l$ = Fact Update$\|D_i, D)$
Sir Anthony Seldon, vice-chancellor of the University of Buckingham, said: "Cheating should be tackled and the problem should not be allowed to fester any longer. "
He added: "This shows the extent to which a party which had such a proud record of fighting racism has been poisoned under Jeremy Corbyn. "
But he said his dream of making it in the game had turned into a nightmare. “
Adam Price, Plaid Cymru leader, said: "There is now no doubt that Wales should be able to hold an independence referendum. "
Others told how excited they had been when they were scouted by Higgins. “
The former Conservative deputy prime minister said it was “complete nonsense” to suggest Brexit could be done by Christmas. “
He said the QAA identified 17,000 academic offences in 2016 - but it was impossible to know how many cases had gone undetected. "
Nationalism leads a "false trail" in ""exactly the opposite direction", he argued, "one that pits working people against each other, based on the accident of geography".
He also suggested that universities should adopt "honour codes", in which students formally commit to not cheating, and also recognise the consequences facing students who are subsequently caught.
He added: "But my experience is, if you make that threat, you don’t actually need to follow through with the dreaded milkshake tax. "
He said: “There’s an anger inside of me, a feeling of disgust that turns my stomach. ”
Damian Hinds says it is "unethical for these companies to profit from this dishonest business".
She added: “His plan to hold another two referendums next year – and all the chaos that will bring – will mean that his government will not have time to focus on the people’s priorities. “
We would be happy to talk to the Department of Education about their concerns." ’
I am determined to beat the cheats who threaten the integrity of our system and am calling on online giants, such as PayPal, to block payments or end the advertisement of these services - it is their moral duty to do so," said Mr Hinds.
The chief executive of Action on Smoking and Health, Deborah Arnott, also warned it would be a "grave error" to move away from taxing cigarettes. "
Rather than just taxing people more, we should look at how effective the so-called ’sin taxes’ really are, and if they actually change behaviour. "
He added: "How many more red lines will be laid down by sensible Labour MPs, only for the leadership to trample right over them?
This shows that the complaints process is a complete sham," she tweeted. "
Mr Hinds added that such firms are "exploiting young people and it is time to stamp them out". "
One said he was abused by Higgins in a gym.

Table 10: Sample of the least likely fact-update sentences, as judged by our best-performing model. Predictions represent a combination of opinion quotes or anecdotes, projects and longer-term plans.