# Revisiting Sentence Union Generation as a Testbed for Text Consolidation

**Eran Hirsch**[1] **Valentina Pyatkin**[1] **Ruben Wolhandler**[1]
**Avi Caciularu**[1] **Asi Shefer**[2] **Ido Dagan**[1]
[1] Bar-Ilan University [2] One AI
eran.hirsch@biu.ac.il dagan@cs.biu.ac.il

## Abstract

Tasks involving text generation based on multiple input texts, such as multi-document summarization, long-form question answering and contemporary dialogue applications, challenge models for their ability to properly *consolidate* partly-overlapping multi-text information. However, these tasks entangle the consolidation phase with the often subjective and ill-defined content selection requirement, impeding proper assessment of models' consolidation capabilities. In this paper, we suggest revisiting the *sentence union* generation task as an effective well-defined testbed for assessing text consolidation capabilities, decoupling the consolidation challenge from subjective content selection. To support research on this task, we present refined annotation methodology and tools for crowdsourcing sentence union, create the largest union dataset to date and provide an analysis of its rich coverage of various consolidation aspects. We then propose a comprehensive evaluation protocol for union generation, including both human and automatic evaluation. Finally, as baselines, we evaluate state-of-the-art language models on the task, along with a detailed analysis of their capacity to address multi-text consolidation challenges and their limitations.[1]

## 1 Introduction

In order to acquire knowledge on a new subject or find answers to complex questions, it is often necessary to consult multiple sources of written information. While information provided in a single document is usually consistent, textual materials from various sources often use different language expressions, which may vary in terms of level of specificity, to convey similar information. An illustration of this phenomenon can be seen in Figure 1. In this paper, we aim to address the process of combining such multiple partially overlapping textual sources into a single unified and comprehensive format, to which we refer as *text consolidation*.

[S1] *The fire has destroyed a large section of the store and fire crews and investigators are still on the scene.*
[S2] *A FIRE has badly damaged the Waitrose supermarket in Wellington's High Street.*

[Union] *The fire has destroyed a large section of the Waitrose supermarket in Wellington's High Street and fire crews and investigators are still on the scene.*

Figure 1: An example of a sentence pair and its union sentence. Information that must be included in the union is highlighted differently for each sentence (*green* and *purple* for sentences 1 and 2, respectively), unless the information is paraphrastic (equivalent) between the two sentences, which is then highlighted by the same color (*blue*). Non-highlighted information indicates that there is corresponding information in the other sentence that is more specific.

Text consolidation plays a crucial role in almost any text-based information access application, such as Multi-Document Summarization (MDS) (Fabbri et al., 2019; Giorgi et al., 2022), long-form question answering (Fan et al., 2019; Nakano et al., 2022), and contemporary dialogue applications (Thoppilan et al., 2022; OpenAI, 2023). It is important to point out here that content selection and consolidation manifest two distinct sub-tasks in such applications, where the former involves identifying the sought information in the source texts, based on considerations such as salience and user needs. Consolidation, on the other hand, involves merging the selected information into a coherent output text. Accordingly, we suggest that each sub-task deserves separate investigation, while focusing in this paper on the consolidation task, manifested as information union. This approach enables targeted investigation of information union capabilities of models, while enabling modular architectures, where an effective information consolidation model can be paired with different content selec-

[1] Our data and code is available at: https://github.com/eranhirs/sentence_union_generation

tion models and strategies, whether fully-automatic or interactively involving a user in the loop.

To achieve a more controlled research environment, a sentence fusion task was introduced, which fuses a set of sentences into a single sentence (Barzilay et al., 1999; Thadani and McKeown, 2013; Agarwal and Chatterjee, 2022). However, being similar to summarization, the general sentence fusion task is ill-defined, because it allows for *subjective* salience-based content selection decisions (Daume III and Marcu, 2004; Krahmer et al., 2008). In contrast, the sentence union generation task is strictly defined as generating a sentence that contains *exactly all* information from the source sentences (see Fig. 1). While identifying the union task to be more attractive due to its more *objective* and semantically challenging nature, we found that datasets for this topic are relatively scarce (McKeown et al., 2010; Geva et al., 2019; Lebanoff et al., 2020), none of them sufficiently addressing the text consolidation setting.

Consequently, we revisit the sentence union generation task and propose that it can be used as an effective generic testbed for text consolidation. Compared to the sentence intersection task, the union task is more challenging, as it requires merging both joint and disjoint information in the output and hence provides a more complete testbed for text consolidation. Our input format is rich and challenging enough, as shown in our analyses, to support research on information merging models. Further, this setting may already be of practical use for downstream text generation tasks, for example when combined with sentence compression or decontextualization models.

Our contributions are outlined as follows: (1) we suggest focusing on sentence union generation as a resource for studying cross-text consolidation capabilities, and point out that properly identifying informational relations between pairs of sentences is necessary for proper consolidation; (2) we provide the largest union fusion dataset to date, while proposing a controlled annotation protocol and interface for careful creation of a sentence union corpus; (3) we suggest evaluation protocols to assess the quality of a generated sentence union, accompanied by automatic metrics that can be used for comparing multiple systems; (4) we provide empirical results on the abilities of prominent neural generative models to address the union task, assessing their capabilities and limitations.

## 2 Background

In Multi-Document Summarization (MDS) (Narayan et al., 2018; Fabbri et al., 2019) multiple-texts are summarized into a single, shorter text. In a more controlled variant of MDS, the task requires the fusion of partly-overlapping sentences (Barzilay et al., 1999; Thadani and McKeown, 2013; Agarwal and Chatterjee, 2022). Generally, the sentence fusion task included a saliency detection (or importance) component which requires identifying which pieces of information to preserve in the fused output. As a result, sentence fusion is generally ill-defined, as different possible content selections may be valid, making the task subjective to varying necessities of a user (Daume III and Marcu, 2004; Krahmer et al., 2008). Its output could be seen as covering a "loose" intersection of the content of two sentences.

McKeown et al. (2010) on the other hand, to ensure more consistent fusion settings, makes a distinction between two strict variants of the task: sentence intersection and sentence union generation. Given two (or a set of source sentences), their intersection is a sentence that contains only information that is *common* to both source sentences, while their union is a sentence that contains *all* information from the source sentences. As we will see in §3, these tasks can indeed be formulated in strict entailment terms. McKeown et al. (2010) crowdsourced a dataset of 300 examples for sentence intersection and sentence union, but subsequent works mostly focused on the intersection fusion part of the dataset (Thadani and McKeown, 2011; Fuad et al., 2019). Further, their dataset size is relatively small and primarily intended for evaluation purposes, making it inadequate for partitioning into a training dataset for fine-tuning large language models.

While McKeown et al. (2010) used similar sentences, whose contents partly overlap, as input, later works researched the union of disparate sentences (Geva et al., 2019; Lebanoff et al., 2021) where contents are disjoint. This does not address the challenge of consolidating partly overlapping texts. In this work, we chose sentence union as a more complete testbed for multi-text consolidation. We see our work as a continuation of the work by McKeown et al. (2010), and complementary to works that introduced fusion datasets for disparate sentences.

Our work further relates to a line of research

that focuses on objective generation of text. Castro Ferreira et al. (2020) introduced a data-to-text generation task, wherein knowledge graph triplets describing facts are transformed into natural language text. While there are many possible realizations of the knowledge graph into natural language, the task is semantically objective, with respect to the informational content expected in the output, and is hence similar to the sentence union task. Recently, Slobodkin et al. (2022) introduced a new *controlled text reduction* task: given an input document with highlighted spans, the task is to generate a summary in which only the information covered in the highlighted spans is included, which could be compared to a highlight union task. Compared to our work, the spans that they used all appear in a single document, which makes it more similar to datasets which fuse disparate sentences.

## 3 Task Formulation

The input for our sentence union task consists of two related sentences whose content partly overlap. The output union is then defined as a single sentence that follows two conditions: (a) it contains exactly the information from the two input sentences, and (b) it does not include any redundancies in its content. Condition (a) implies that there cannot be any information missing from the union that is mentioned in the source sentences, while at the same time the union cannot contain information that is not mentioned in the source sentences (i.e., hallucinations). Condition (b) implies that the union must avoid repetition of any units of information stemming from the source sentences, even if they are conveyed in different lexical terms.

Notably, the semantic content of the output union (condition (a)) can be defined objectively in strict textual entailment terms. Formally, given an input of two related sentences $s_1$ and $s_2$, and their union $u$, $u$ should satisfy $u \models s_1$ , $u \models s_2$ and $s_1 + s_2 \models u$, where $\models$ denotes textual entailment and $+$ denotes concatenation of the two sentences. This definition, however, does not cover condition (b) of avoiding redundancies.

Identifying relevant informational links is crucial for producing a union, as demonstrated by the example in Fig. 2. We observe three types of relations between information units in the source sentences that affect the content of the resulting unit: (1) equivalent content, (2) uni-directional entailing content, and (3) disjoint content. Equivalent content, such as lexical equivalence or paraphrases, needs to be identified and included exactly once in the union to avoid redundancy. Uni-directional entailing content pertains to aligned text spans where one span can be implied from the other. In this case, only the entailing text unit should be included: including both spans would be redundant, while including only the less specific mention would result in missing information. Disjoint content must be included in the union as it provides distinct information not mentioned in the other sentence. For example, in Fig.2, sentence 1 mentions the reason for firing Weightman while sentence 2 mentions that Harvey resigned, each providing distinct information. In addition, according to our annotation scheme, we assume that the date of the publication is known, which means that when a phrase such as "the previous Thursday" is mentioned, we can infer the specific date. Thus, the text spans "On March 1st" and "the previous Thursday" are equivalent, while "Francis Harvey" in sentence 1 is more specific than the text span "Harvey" in sentence 2. By considering these three types of relations, a proper union can be produced.

As noted earlier, we see the union generation task as a more comprehensive setup for information consolidation than the *intersection* generation task[2]. This is because the union output should combine all the content from both source sentences, while the output of the intersection task does not include information mentioned in only one of the sentences. As a result, the union is more informative than the intersection, which makes it more representative for downstream multi-text tasks requiring information consolidation, aiming to create an efficient, non-repetitive output text.

## 4 Dataset

### 4.1 Data sources

Annotating a text consolidation sentence union dataset requires a collection of *related* sentences, as input, as seen in Fig. 1. Specifically, we require naturally occurring sentences with some semantic overlap, where different types of informational relations are present. Note that we do not consider sentences with no content overlap as relevant for our dataset.

[2]The information content for the intersection task can also be defined in strict textual entailment terms. Formally, for the intersection $i$ of the two sentences $s_1$ and $s_2$, it is required that $s_1 \models i$ , $s_2 \models i$ and for all $i^*$ such that $s_1 \models i^*$ , $s_2 \models i^*$ , then $i \models i^*$.

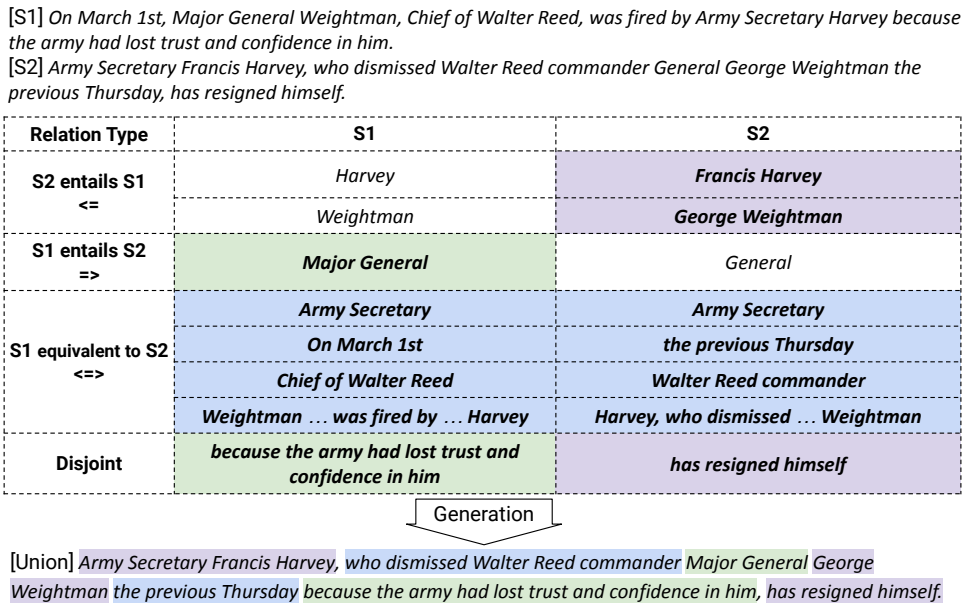



Figure 2: An example of a pair of sentences, the informational relations between their text spans, and their union. In order to generate the union, it is first necessary to identify these relations (possibly implicitly), and then include all new or more specific information (denoted by colors) without redundancy.

To that end, we use the dataset created by Weiss et al. (2021), which includes pairs of relevant sentences with high semantic overlap. Their dataset was curated by identifying information overlap between sentences, based on the repurposing of existing human annotations. This approach is preferable to using models that identify semantic overlap, such as Thadani and McKeown (2013), since it introduces less bias to the dataset. The original datasets from which they sourced the sentences include: (1) the Event Coreference Bank (ECB+, an extension over ECB) (Cybulska and Vossen, 2014), which provides annotations for coreferring event and entity mentions, (2) MultiNews (MN) (Fabbri et al., 2019), which contains clusters of news articles along with human-written summaries, and (3) The Document Understanding Conference (DUC) and the Text Analysis Conference (TAC)[3], both providing MDS evaluation datasets.

## 4.2 Annotating sentence union

The process of writing a sentence union involves carefully tracking information units and blending them together to form the output, as outlined in §3. We introduce an elaborate crowdsourcing approach and interface (see Figure 3) for annotating union datasets at a large scale, which splits the annotation process into multiple steps.

Starting with the two source sentences, the first step is to choose one sentence as the *base sentence*,

[3] https://duc.nist.gov/ , https://tac.nist.gov/

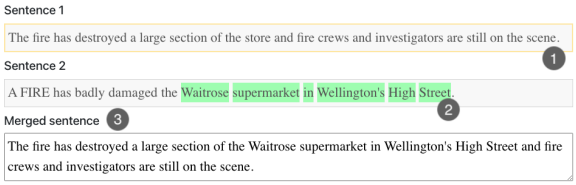



Figure 3: A screenshot of the sentence union text generation annotation interface. The screenshot shows the last step, where the worker already choose sentence 1 as the base sentence [1], highlighted the new or more specific information in sentence 2 [2] and wrote the final sentence union ("Merged sentence") [3].

that will be used as the basis for generating the sentence union, depicted in (Fig. 3, [1]). Our early experiments have shown that it is easier to merge the information from one sentence by adding it to the other sentence than write a merged sentence from scratch. We instruct the workers to choose the more detailed sentence as the base sentence, since this sentence would usually require less edits when merging into it information from the other sentence. In the other sentence, termed the *integrated sentence*, the worker has to highlight which spans they would like to integrate into the base sentence (Fig. 3, [2]). Finally, in the writing step, the worker blends the highlighted spans into the base sentence, thus creating the sentence union (Fig. 3, [3]).

To optimize the diversity of inputs within our dataset while considering our annotation budget, each example was assigned to a single annotator.

| Split | Train | Dev | Test | Skipped |
|---|---|---|---|---|
| Size | 1087 | 349 | 477 | 458 |

Table 1: Sizes of the splits of our dataset, as well as of the skipped examples (19.3% of Weiss et al. (2021)).

| Datasets | Coverage | Faithfulness | Redundancy |
|---|---|---|---|
| Ours | 98.3% | 99.8% | 99.8% |
| McKeown et al. (2010) | 96.5% | 99.5% | 98.6% |

Table 2: Evaluation of union quality.

To ensure the quality in annotators' decisions, our process follows the controlled crowdsourcing approach (Roit et al., 2020). See App. C for more details and screenshots of the entire annotation process.

**Skipping examples** In certain cases, it may not be possible to generate a coherent sentence union from a pair of sentences, and annotators were given the option to skip such examples. A comprehensive analysis of these skipped cases is presented in Appendix A. Mainly, our findings indicate that the dataset from which we derived our data(Weiss et al., 2021), and was primarily designed for proposition alignment, contains many sentence pairs that are not sufficiently related to each other and hence are not suitable for producing a meaningful union.

**Subtle annotation cases** In addition to the aforementioned instructions, we took into consideration a few prominent special cases concerning the source sentences that would affect the resulting sentence union. Such cases include the need for world knowledge, temporal issues, subjectivity and attribution. For examples and guidelines provided to the workers for such cases, refer to App. B.

### 4.3 Cleaning annotations

In order to ensure a high quality dataset, we introduced a post-processing step in which we either removed or manually edited examples matching specific filtering criteria. Filtering included finding non-overlapping input sentences based on their output union (i.e., the output was a simple concatenation of the two source sentences), as well as automatically identifying and manually reviewing subtle annotation cases described in App. B. For more details, see App. D.

## 5 Dataset Analysis and Assessment

In the following subsections, we report various analyses of the quality and other properties of our dataset. Dataset split statistics appear in Table 1. Our approach yielded a test dataset comprising of 477 instances, a sample size which is reasonable in light of the confidence intervals outlined in §8. Moreover, our analysis of learning curves (see Appendix G) suggests that the size of our training dataset is sufficient, and further expansion may not yield significant benefits.

### 5.1 Sentence union quality

To estimate the reliability of our dataset, we have conducted a human assessment on a sample of 100 examples of sentence unions generated by our annotators. Our goal is to check whether the sentences in the dataset objectively fulfill the union requirements defined in Sec. 3. For this purpose we designed two evaluation criteria for content (*coverage*, *faithfulness*), and one criterion for finding redundancies (*redundancy*). In addition, we evaluate the fluency of the generated sentence, as commonly done for generation tasks.

- **Coverage:** Does the sentence union contain *all* information expressed in the source sentences?
- **Faithfulness:** Does the sentence union describe *only* information expressed in the source sentences?
- **Redundancy:** Does the sentence union redundantly repeat some information?
- **Fluency:** Does the sentence union progresses fluently, form a coherent whole and is easy to understand?

The content criteria resemble closely those used for data-to-text generation tasks (Castro Ferreira et al., 2020) which also require exact content matching between their input and output. We add another criterion for evaluating redundancies, as our input does include redundancies which needs to be avoided in the output.

As a simple way to measure the content criteria, we count the number of content words[4] involved in pieces of information that are missing from the sentence union, or are unfaithful to the source sentences. For example, if the sentence union in Fig 2 would not mention the name *"Nick Jones"*, which was mentioned in sentence 2, we count this as 2

[4]We removed stop words using `www.nltk.org`.

misses. A more complicated example would be if the sentence union attributes *"Nick Jones"* to the wrong entity, such as *"FBI Deputy Director Nick Jones"*. In such case, we consider the entire span (5 words) as missing, as well as unfaithful. Note that faithfulness can be seen as symmetrical to coverage, where we simply count content words in the sentence union that are not supported in the source sentences. Similarly, for the redundancy score, we count the number of content words involved in pieces of information that are redundant in the union. For example, in the phrase *"Thursday overnight at 2:09am"*, the phrase *"overnight"* is considered redundant, and we will count 1 redundant word. We did not notice any fluency issues in the sentence unions created by the workers, as may be naturally expected given the high quality of our selected workers.

We start by counting the number of content words in all of the sentence unions in our sample, which adds up to 2372 content words, termed $w_{total}$. Then, to create a *coverage* score, the count of missing content words is termed $w_{missing}$, and the coverage score is calculated as $\frac{w_{total}}{w_{total}+w_{missing}}$. To create a *faithfulness* and *redundancy* scores, we calculate $1-\frac{w_{unfaithful}}{w_{total}}$ and $1-\frac{w_{redundant}}{w_{total}}$, respectively, where $w_{unfaithful}$ is the number of unfaithful words and $w_{redundant}$ is the number of redundant words. Results for these metrics are available in Table 2. Overall, coverage issues were encountered in 8 examples out of 100, faithfulness and redundancy issues in one example each.

**Quality comparison to the prior dataset** We compare our dataset to the McKeown et al. (2010) dataset of 300 sentence unions examples. In their annotation process, 5 workers annotated each pair of sentences, and then a single sentence union out of the 5 was automatically chosen as a representative. We evaluated a sample of 20 such representative sentence unions and used the same quality metrics that were used in our dataset quality analysis, reported in Table 2. We conclude that our controlled process, which separates the identification of informational relations from the writing phase, results in higher quality sentence unions, making significantly less coverage and redundancy mistakes, which are often due to lack of attention to details. For the faithfulness criterion, both approaches achieved similar high scores, which is expected since humans are not prone to hallucinate when editing a sentence. Overall, our annotation process achieves slightly better results, while employing only one worker instead of five.

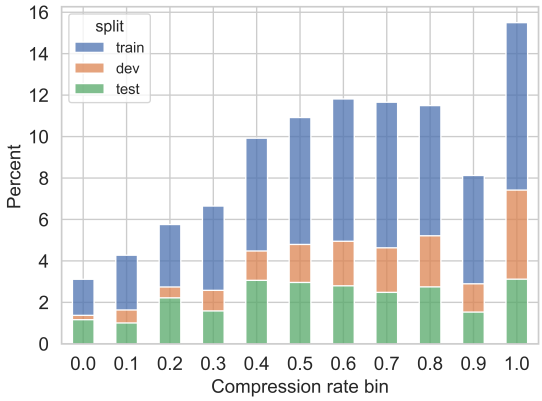



Figure 4: Compression Rate (CR) vs. the frequency of each CR bin, for the train/dev/test dataet splits.

## 5.2 Dataset compression rate

Our motivation for the union task is to develop models that can consolidate information from naturally occurring texts with varying degrees of overlapping information. Hence, in order to assess the diversity of our dataset with respect to the degree of such information overlap, we suggest to compute and analyze the *Compression Rate* (CR) in our instances, which measures in our setting the amount of redundancies (unlike the data-to-text setting) between the two source sentences[5]. By design, a CR of 100% would imply that a single source sentence contains all of the information in both source sentences, which means that the other sentence is completely redundant. A CR of 0% would imply that there is no redundancies between the source sentences.

Denoting our two input sentences `short` and `long`, per their lengths, as well as the `union` sentence, and following the rationale above, the compression rate is calculated as the amount of information that is eliminated from the shorter sentence. Formally, we have $\text{CR}(\texttt{short}, \texttt{long}, \texttt{union}) = 1-\frac{|\texttt{union}|-|\texttt{long}|}{|\texttt{short}|}$, counting sentence length by content words.

As can be seen in Fig. 4, our dataset supplies a variety of examples in terms of CR for every split. We report an average CR score of $60.82_{\pm 0.67}$ for our dataset and an average CR score of $65.62_{\pm 1.35}$ for McKeown et al. (2010). These results imply that our dataset on average contains somewhat less

[5] In the union task, compression refers only to the merging of redundancies across the source sentences.

overlap between the source sentences, overall includes a large variety of redundancy levels.

### 5.3 Informational relations analysis

Complementary to the analysis in §5.2, naturally occurring texts can include a wide variety of cross-text informational relations, as described in §3. For this reason, we analyzed the frequency of the more challenging relations necessary to generate proper sentence union. Our analysis includes a sample of 30 sentence pairs from our dataset. On average, a sample of 10 examples is expected to include 17 "paraphrastic uni-directional entailment" relations (a uni-directional entailment which differs lexically), such as *"supermarket"* entailing *"store"*, or *"gave interviews on NBC's today"* entailing *"appearance on NBC's today"*. As described in §3, such examples challenge a consolidation model to include only the *entailing* expression in the output. In addition, such a sample is expected to include 21 paraphrastic equivalence relations. These challenge the model to include only one of the equivalent expressions in the output, to avoid repetition. Overall, these statistics assess the abundant semantic challenges posed by our dataset.

## 6 Baseline Models

We present baseline models, aiming to test neural pretrained language models' for their ability to implicitly recognize relevant informational relations between input sentences and properly create their union.

**Fine-tuned models** As our first type of baseline we fine-tune a large pre-trained sequence-to-sequence model using our data. To that end, we picked two strong models: $T5_{large}$ (Raffel et al., 2019), which is commonly applied to end-to-end text generation tasks (Chen et al., 2020), and PRIMERA (Xiao et al., 2022), which was pre-trained in a cross-document fashion (Caciularu et al., 2021) and achieves state-of-the-art results over multi-document summarization datasets. This makes this model appealing for our sentence fusion task, where the two sentences originate in different documents. See App. F for information about training details.

**In-context learning** Another current baseline approach is in-context learning, in which the instructions and examples to the task are provided as input (the prompt) at inference time to very large pre-trained language models. We used $GPT3$ (Brown et al., 2020), specifically *text-davinci-003*. The instructions we initially used were similar to those given to the annotators. We then optimized the prompt by running it on the training dataset and manually identifying mistakes. The identified mistakes were added to the prompt as examples. In addition, we added to the instructions "important" notes to what the model should pay attention to. See App. E for the complete final prompt and configuration used.

| Score | Content | Redundancy |
|---|---|---|
| 1 | Substantial information is missing. | Substantial information is repeated. |
| 2 | Some information is missing. | Some information is repeated. |
| 3 | Minor details are missing. | Minor details are repeated. |
| 4 | Nothing is missing. | Nothing is repeated. |

Table 3: The ordinal scales used for the content (coverage & faithfulness) and redundancy measures.

## 7 Model Evaluation Protocols

We evaluate our baseline systems both through human evaluation (§7.1) and with automatic metrics (§7.2) suitable for the task, which can generally be used in the development cycles of union generation systems (§7.2).

### 7.1 Human evaluation

The human evaluation is conducted over the predicted unions for the test set for each of the baseline models. Instead of judging the generated sentence union for each baseline system separately, the evaluation is done in a comparative fashion, following previous works where the evaluator sees together the outputs of all baseline systems (Callison-Burch et al., 2007; Novikova et al., 2018).

Similar to the analysis of the dataset quality in §5, we are interested in evaluating the coverage, faithfulness, redundancy and fluency of the predicted union, this time in a manner that fits crowd-sourced human evaluation. Content and redundancy are scored on a scale from 1 to 4 (higher is better), described in Table 3. This scale is inspired by the Semantic Textual Similarity human evaluation approach (Agirre et al., 2013), which also tests for information overlap. For the fluency score, we use a common Likert scale from 1 to 5 (Fabbri et al., 2021). See App. H for details and screenshots.

As there exist trade-offs between the two content measures and the redundancy measure, we add an additional measure which evaluates *consolidation*

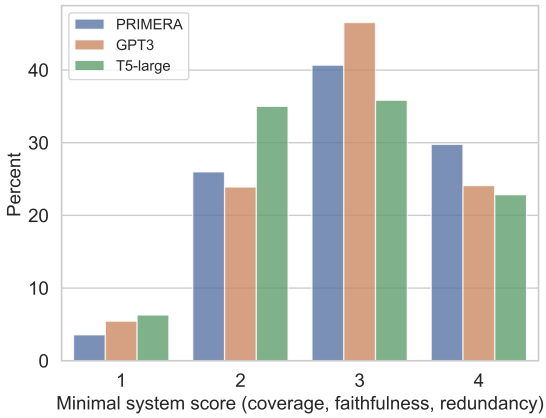



Figure 5: A histogram of minimal system scores, testing for coverage, faithfulness or redundancy mistakes.

as a whole. For example, by arbitrarily adding more information to the union we can increase the coverage, but also risk increasing redundancies and unfaithfulness. The *consolidation* measure simply averages the three aforementioned measures, thus testing for overall text consolidation quality.

### 7.2 Automatic evaluation

In line with previous works in text generation, we report the ROUGE metric between the reference union and the predicted union. However, like for most generation tasks, ROUGE will unfairly penalize correct but paraphrastic sentence unions (as described in §3). To partly address this issue, we add another automated metric which tests for bi-directional textual entailment (aka NLI), comparing the reference union sentence to the predicted union sentence, requiring entailment in both directions. Specifically, we use the $DeBERTa_{xxlarge}v2$ model (He et al., 2020), fine-tuned with the MNLI task (Williams et al., 2017) and a threshold of 0.5.

While both metrics test for content matching, they would not penalize a model that bluntly concatenates the two input sentences. Therefore, we also report $\Delta CR$ (§5.2), calculated as the average difference between the CRs of the predicted vs. the reference union sentences (the latter is subtracted from the former), on each instance. A positive value thus indicates that the model compression rate is higher than that of the reference union, while a negative value indicates the opposite (model compresses less than the reference).

## 8 Results and Analysis

### 8.1 Human evaluation of the models

Results are presented in Table 4, and example generations with their respective scores are provided in App. I. The trade-off mentioned in §7.1 between increasing coverage while still remaining faithful and without redundancies is evident in the results of $T5_{large}$ and $GPT3$. PRIMERA comes out as a slightly better model, as it achieves the highest consolidation score, with yet a lot of room for improvement.

To get a better sense of the absolute performance of the union sentences generated by the baseline models, we compare them to two naive models which output: (1) the concatenation of the source sentences (no avoidance of *redundancy*), and (2) the longer sentence (no attempt to consolidate and *cover* information from the other sentence). Based on evaluation of 50 examples completed by the authors, we report an average redundancy score of $1.6_{\pm.1}$ for the concatenation and an average coverage score of $2.3_{\pm.1}$ for the longer sentence. As reported below, all our baseline models outperform these naive models by a large margin.

Further, we draw a plot (Fig. 5) of the minimal system score amongst the three component measures that the consolidation measure combines. We note that even for the best model, PRIMERA, only 29.7% of the predictions are fully correct with respect to content and redundancy, another 40.6% examples include minor errors, and 26% examples contain substantial errors in at least one of the measures, indicating the limitations of current models.

### 8.2 Automatic evaluation of the models

While automatic metrics are clearly less reliable than human metrics, they can be useful for development cycles. The automatic metric results are also reported in Table 4, observing that both the $ROUGE1$ score is highest for PRIMERA, while the NLI score is highest for $GPT3$. The $\Delta CR$ scores roughly correlate with the combination of coverage and redundancy detected in the human evaluation, where both lower coverage (undesired) and lower redundancy (desired) increase compression rate.

To identify the potential utility of our automatic metrics, we follow the standard practice (Fabbri et al., 2021) and calculate a Kendall $\tau$ coefficient (McLeod, 2005) between the human and automatic evaluation results. Our results show that $ROUGE1$

| | Coverage (1 to 4) | Faithfulness (1 to 4) | Redundancy (1 to 4) | Consolidation (1 to 4) | Fluency (1 to 5) | $ROUGE1$ | NLI | $\Delta CR$ |
|---|---|---|---|---|---|---|---|---|
| PRIMERA | 3.2 | 3.7 | **3.8** | **3.6** | 4.1 | **89.92**$_{\pm.4}$ | 86.37$_{\pm1.6}$ | 9.28 $_{\pm1.5}$ |
| $GPT3$ | **3.5** | 3.5 | 3.5 | 3.5 | 3.8 | 85.35$_{\pm.4}$ | **96.23**$_{\pm.9}$ | -8.83$_{\pm1.6}$ |
| $T5_{large}$ | 2.8 | **3.8** | **3.8** | 3.5 | **4.2** | 85.88$_{\pm.5}$ | 73.38$_{\pm2.0}$ | 27.2 $_{\pm1.7}$ |

Table 4: Human (left) and automatic (right) evaluation results of system generated unions over the complete test set. All scores are averages, along with their standard error (standard error for manual evaluation results was always smaller than 0.01, and is therefore omitted from the table).

is the highest correlated metric with the consolidation measure ($\tau = 0.38$, $p < 0.05$). Overall, these automatic metrics can be used in tandem to provide certain feedback during model development cycles.

### 8.3 Error analysis

To shed light on the various errors made by the baseline models, we examined 20 erroneous examples identified in the human evaluation, with each example consisting of three predictions, one from each of the baseline systems. Our findings indicate that the most frequent causes of model errors are related to the complexity of informational relationships present in the source sentences, with uni-directional entailment being the most common. Moreover, the models seem to face difficulties in accurately combining related information, which often results in incorrect merging of information with the wrong entity or predicate. Further details on the analysis can be found in Appendix J.

## 9 Conclusions

In this paper, we advocate for using the sentence union task as a testbed for multi-text consolidation. We release a realistic dataset, together with a set of analyses that show that the dataset is of high quality, and challenging for multi-document consolidation efforts. We evaluate the performance of state-of-the-art pretrained large language models on text consolidation, where our findings suggest key challenges for future research.

Future research may expand upon our dataset to include consolidation beyond 2 input sentences, and may examine the use of explicit text consolidation structures for improving multi-text consolidation in large language models.

## Limitations

We enumerate some limitations to our work. While we did create the largest union dataset to date, it is still of moderate size. As shown by our learning curves (App. G), the amount of training data we created seemed sufficient to saturate the learning of the models with which we experimented, but it might still be found insufficient for training other models.

Our annotation protocol might have influenced the compression rates of the unions, as we instructed workers to annotate sentence unions by first choosing a base sentence and then highlighting the other sentence. Additionally, while the highlighting facilitates the annotation process, it cannot directly be used for analyses of the dataset since it is uni-directional.

The dataset includes only input with exactly two sentences and it might be desirable for future works to also be able to train systems that take more than two sentences as input. Our dataset is also domain specific, in that all the sentences are taken from news sources. This might result in challenging cross-domain generalization.

This dataset is limited to the English language. While the suggested annotation protocol seemingly fits other languages, the step in which words are highlighted might prove problematic for morphologically rich languages, in which a single word includes many pieces of information. A segmentation of the text before annotation might be required.

## Ethics Statement

**Crowdsourcing** To crowdsource the dataset, we used the Amazon Mechanical Turk[6] (MTurk) platform. To participate in the first stage of recruitment, workers were required to possess the following MTurk qualifications:

- NumberHITsApproved greater than 10000
- PercentAssignmentsApproved greater than 98%
- WorkerLocale in US, CA, AU, GB, NZ

[6] https://worker.mturk.com/

Workers were paid $0.3 for each sentence union annotation assignment, as well as a $1.25 bonus for every 100 assignments, and $0.4 for each evaluation assignment, as well as a $1 bonus for every 50 assignments. Overall, by an average approximation of 1.8 minutes for the first assignment, and 2.4 minutes for the second assignment, their wage is expected to start from $10 per hour and increase as the workers are more familiar with the task and start receiving bonuses.

Workers were informed that the ratings they will provide will be used to evaluate artificial intelligence models which were trained on the data they annotated.

**Dataset** The texts that workers write that are included in our dataset are limited to the information expressed in the source sentences. The source sentences originate from the datasets mentioned in §4.1, which include only texts available in public news sources and were previously made available by Weiss et al. (2021). Our dataset does not contain information that would make it possible to reconstruct the original documents, or any human annotations, such as the summary or coreference resolution annotation, from the original datasets.

## Acknowledgments

The work described herein was supported in part by grants from One AI, the Israel Science Foundation 2827/21 and the Israel Ministry of Science and Technology. We would like to thank the workers who have annotated this dataset and we appreciate their dedication in ensuring a high level of quality. We express our gratitude to Dr. Kapil Thadani for assisting us in retrieving his data from an earlier research endeavor.

## References

Raksha Agarwal and Niladri Chatterjee. 2022. Improvements in multi-document abstractive summarization using multi sentence compression with word graph and node alignment. *Expert Systems with Applications*, 190:116154.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557, College Park, Maryland, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. 2020. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems (NeurIPS)*.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hal Daume III and Daniel Marcu. 2004. Generic sentence fusion is an ill-defined summarization task. In *Text Summarization Branches Out*, pages 96–103, Barcelona, Spain. Association for Computational Linguistics.

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

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Tanvir Ahmed Fuad, Mir Tafseer Nayeem, Asif Mahmud, and Yllias Chali. 2019. Neural sentence fusion for diversity driven abstractive multi-document summarization. *Computer Speech & Language*, 58:216–230.

Mor Geva, Eric Malmi, Idan Szpektor, and Jonathan Berant. 2019. DiscoFuse: A large-scale dataset for discourse-based sentence fusion. 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 3443–3455, Minneapolis, Minnesota. Association for Computational Linguistics.

John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Lu Wang, and Arman Cohan. 2022. Exploring the challenges of open domain multi-document summarization.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Emiel Krahmer, Erwin Marsi, and Paul van Pelt. 2008. Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion. pages 193–196.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Lidan Wang, Walter Chang, and Fei Liu. 2020. Understanding points of correspondence between sentences for abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 191–198, Online. Association for Computational Linguistics.

Logan Lebanoff, Bingqing Wang, Zhe Feng, and Fei Liu. 2021. Modeling endorsement for multi-document abstractive summarization. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 119–130, Online and in Dominican Republic. Association for Computational Linguistics.

Kathleen McKeown, Sara Rosenthal, Kapil Thadani, and Coleman Moore. 2010. Time-efficient creation of an accurate sentence fusion corpus. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–320, Los Angeles, California. Association for Computational Linguistics.

A Ian McLeod. 2005. Kendall rank correlation and mann-kendall trend test. *R Package Kendall*, 602:1–10.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback.

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

Jekaterina Novikova, Ondř ej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. Association for Computational Linguistics.

Aviv Slobodkin, Paul Roit, Eran Hirsch, Ori Ernst, and Ido Dagan. 2022. Controlled text reduction.

Kapil Thadani and Kathleen McKeown. 2011. Towards strict sentence intersection: Decoding and evaluation strategies. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 43–53, Portland, Oregon. Association for Computational Linguistics.

Kapil Thadani and Kathleen McKeown. 2013. Supervised sentence fusion with single-stage inference. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1410–1418, Nagoya, Japan. Asian Federation of Natural Language Processing.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng,

Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Daniela Brook Weiss, Paul Roit, Ayal Klein, Ori Ernst, and Ido Dagan. 2021. Qa-align: Representing cross-text content overlap by aligning question-answer propositions.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

## A Skip Guidelines

In Section 4.2, it was noted that there are cases where generating a union from a pair of sentences is not suitable, and workers were given the option to skip the annotation for such examples. This section outlines the specific scenarios in which workers were directed to skip examples. Eventually, our annotators skipped 458 sentence pairs from the original dataset that we used as input, as shown in Table 1. An analysis of a sample of 30 such cases is presented in Table 5, categorized based on the criteria below. In conclusion, we found that the dataset we used as the source of our sentence pair instances, which was originally developed by Weiss et al. (2021) for aligning predicate-argument structures (represented as question-answer pairs), includes a significant number of instances where information consolidation in the form of sentence union is mostly irrelevant.

| Category | Count |
|---|---|
| No information consolidation | 19 |
| Unnatural union | 7 |
| Mistake | 3 |
| Missing context | 1 |

Table 5: An analysis of 30 cases that were skipped by workers during the annotation process. Among these, some were categorized as mistakes, meaning that they should not have been skipped.

**No information consolidation.** One case in which workers were directed to skip examples during annotation is when there is no partially overlapping information to consolidate from two related sentences, hence their union would simply be a concatenation of the two. This case is referred to as "No information consolidation". An example of this scenario is when sentence 1 mentions that *"Acupuncture is the ancient Chinese medical therapy technique of inserting thin, sharpened needles into specific nerve junction points of the body,"* and sentence 2 mentions a study that found *"53.8 percent of the subjects who had needles inserted in four acupuncture "zones" in the ear five times a week tested free of cocaine at the end of the eight-week study period."* In this case, there is no need to consolidate the information from the two sentences as they provide distinct pieces of information. Sentence 1 explains what is acupuncture while sentence 2 discusses a study about it.

**Unnatural union.** An example of an "Unnatural union" scenario is when unifying two input sentences would form an awkward or unnatural sentence. For instance, if the first sentence is written in the past tense and the second one in the future tense, unifying them could lead to an unnatural sentence union. As an example, consider the following sentences: *"Fannie Mae's board met Sunday night to discuss Raines' future"* and *"The directors of Fannie Mae, the big mortgage finance company, will meet Sunday to consider the fate of two senior executives who signed off on financial statements that violated accounting rules, people close to the company said Friday."* Here, the first sentence uses

the past tense while the second sentence uses the future tense. It would be more natural to use the past tense in the sentence union since the event occurred in the past. However, incorporating the information that someone said something on Friday before the event could result in an awkward sentence union.

**Missing context.** This case happens when two sentences need to be interpreted in the broader text context, which is missing in our annotation scenario, for example when there is a dangling reference to an entity that is not specified in the given sentence. This is often not problematic, unless understanding the identity of the entity is necessary to create the union. For instance, one sentence quotes a person, while the other sentence does not mention the speaker. An example of this scenario is the following: *"Sadly, because Magic Leap seldom hires and does not actively recruit female candidates, the company loses competitive advantage to products like Microsoft's Hololens."* and *"When Tannen Campbell was hired by Magic Leap in 2015, the Florida company had no women in leadership roles and its only idea to make its product female-friendly was to release a pink version, according to Forbes."* Merging these two sentences is not straightforward due to the lack of context.

**Disagreements.** Sometimes, there are two statements that contradict or disagree with one another. For example, sentence 1 is *"Video of Brooklyn Mother of 13 Zurana Horton shot and killed in a gang shooting was revealed Thursday ."* and sentence 2 is *"A shocking video released for the first time Thursday captures the moment a Brooklyn mother of 12 was killed in a gang shootout as she picked her daughter up from school ."*. Sentence 1 mentions that the child is 13 years old while sentence 2 mentions that the child is 12 years old.

| Category | Count |
|---|---|
| Attribution | 12 |
| Relative dates | 4 |
| World knowledge | 2 |
| Before and after an event | 0 |
| No subtle case of above categories | 34 |

Table 6: Distribution of subtle annotation cases in a sample of 50 instances (some instances belong to more than one category).

## B Subtle annotation Cases

In Section 4.2 we noted that certain special cases arose when generating a union from a pair of sentences, and were included in the instructions for annotators. This section outlines the specific instructions provided to workers, with an analysis of 50 cases (Table 6), categorized based on various criteria as described below.

**Attribution.** One potential issue is when the source sentences make attributions to a specific source, such as a news agency. An example of this can be seen in sentence 1 *"Video of Brooklyn Mother Zurana Horton being shot and killed was revealed Thursday, according to the N.Y. Daily News."* and sentence 2 *"A shocking video released for the first time Thursday captures the moment a Brooklyn mother was killed as she picked her daughter up from school."*, where the new information in sentence 2 is attributed to the video content, rather than to the N.Y. Daily News. Another example is when a sentence contains quotes, as changing a quote to contain more information would create an unfaithful sentence union. In such cases, the workers were allowed, whenever it seemed reasonable, to attribute combined pieces of information originating from the two sentences to a reported source, even if only parts of the combined information were explicitly attributed to this source, in one of the sentences.

**Relative dates.** Some sentences may mention a specific time relative to when the sentence was written, such as "yesterday" or "Monday", which implies that the sentence was written in the same week of the event. Workers were instructed to assume that the date of publication is known, so there is no difference between the mention of "yesterday" and "Monday", but, for example, that "yesterday" is more specific than "earlier this month".

**World knowledge.** In some cases, sentences may mention the same piece of information in different levels of specificity, which requires world knowledge to identify. Workers were instructed to assume common world knowledge when creating the sentence union. An example is given for Paris, which is both a city in Texas and the capital of France.

**Before and after an event.** For sentences referring to events, some may differ in their time of publication compared to the event itself. Workers were instructed to use the past tense, as the

sentence union is written after the event. For example, sentence 1 mentions an event that has already happened *"After leaving Alderson at 12:30 a.m. on March 3, 2005, Martha Steward declared the 5-month experience as "life altering and life affirming.""*, while sentence 2 was written before the event *"US lifestyle guru Martha Stewart is expected to leave jail on Friday after a five-month sentence for a stock scandal that reinvigorated her career rather than dooming it."*. In this case, the sentence union should be written in the past tense, as it refers to an event that has already occurred.

## C Annotation Process

Screenshots of the entire annotation process are depicted in Figure 6. Guidelines for creating sentence unions[7] include writing one coherent sentence, ordering the information in a stand-alone manner (as if the sentence would have been written from scratch), meaning that the writing process should not be distracted by the original split and ordering of information in the two input sentences. To the extent possible, the sentence union should preserve the original wording of the information, but phrasing may be *minimally* adjusted to create a coherent sentence union. Each piece of information should appear only once in the sentence union. When there is a redundancy across the two sentences, the more specific phrasing should be chosen.

The interface helps the workers to avoid making common mistakes. For example, in order to reduce redundancies of information in the union, if a highlighted word already exists in the base sentence, both word mentions will be marked to draw the worker's attention. Another example is warning the worker when the sentence union contains non-highlighted words from the base sentence. Also, when integrating highlighted words into the sentence union, the worker will see yellow highlights turn into green highlights. If the worker tries to submit the annotation with yellow highlights, the system will raise an alert.

To ensure the quality in annotators' judgements, our process follows the controlled crowdsourcing approach (Roit et al., 2020), which includes a recruitment phase, two training phases accompanied by extensive guidelines, and ongoing monitoring during the annotation of the production task. Workers were allowed to participate in primary tasks only if they had completed the entire process. Only workers who performed well on the recruitment phase were accepted to the next training phases. The training phases were created manually, including subtle annotation cases. After each annotation, workers were shown gold target highlights and sentence unions[8] for comparison with their own output.

## D Cleaning Annotations

**Disjoint sentences** Following the skip guidelines (see App. A), we automatically identified examples which their sentences are mutually exclusive and their sentence union is a concatenation of the source sentences. We find these instances by comparing content words only, since connecting the two sentences sometimes involves non-semantic lexical changes (e.g., adding a semicolon or a comma). Due to the fact that there is no consolidation of information in such examples, we see them unfit for a union, as mentioned in §4.1, and they were not included in the dataset. We leave the automatic categorization of sentences into whether or not they are suitable for sentence unions to future work.

**Quotes** Following the attribution discussion in App. B, we manually reviewed examples where the union contained a quote that was not in any of the source sentences, as well as any example that had a sentence which used a first-person perspective (e.g., "I", "we", "mine", "ours", ...).

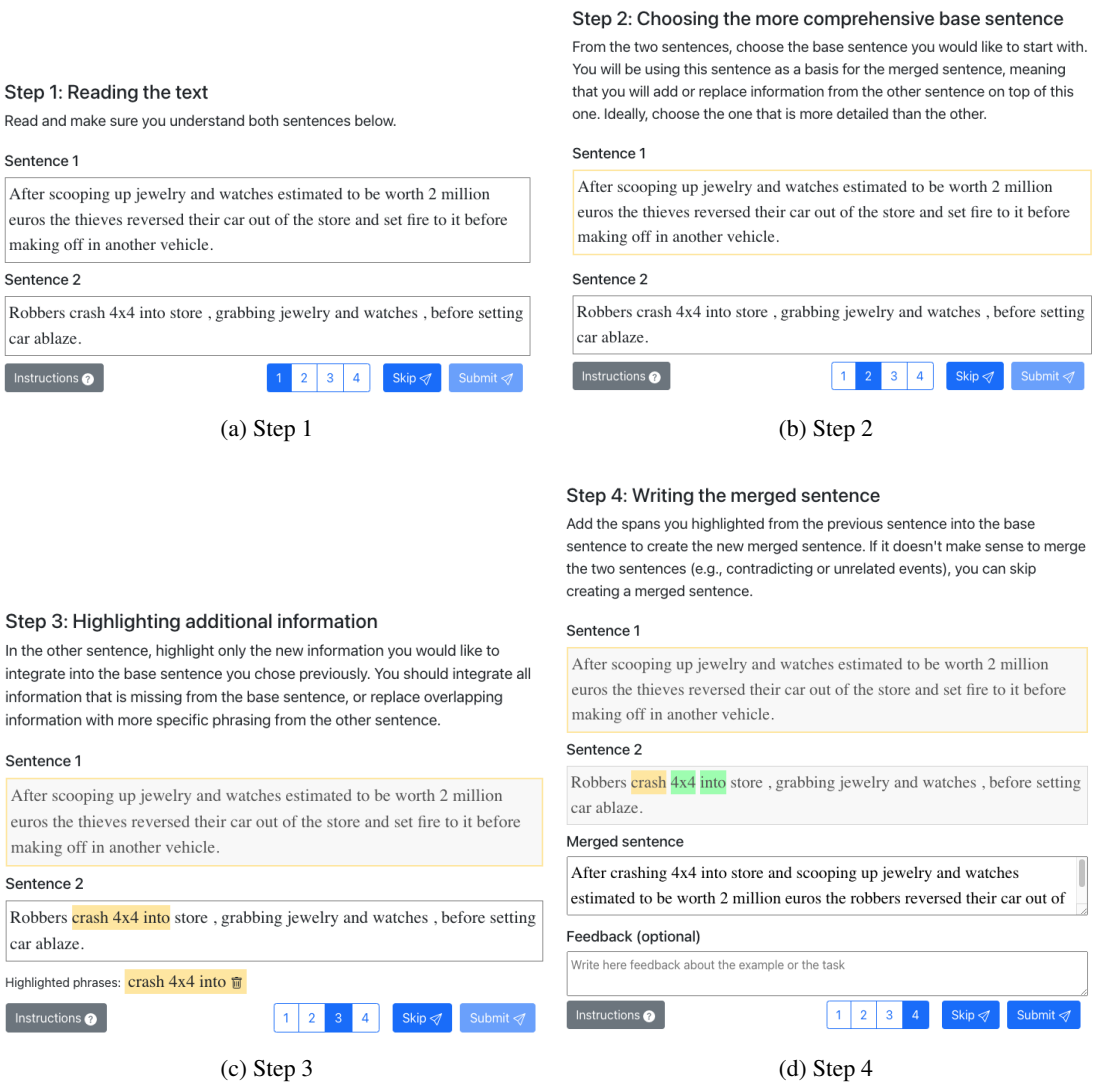
## E In-Context Learning

For the in-context learning approach, we used a temperature value of 0.4 and the following prompt: *In this task, you will be presented with two sentences that overlap in information, and you are tasked to merge the information of the two into a single unifying sentence without redundancies. Important: Do not omit information. Important: Do not repeat information.*

*Here is an example of a correct union and a wrong union: Sentence 1: The February assassination of former Lebanon Prime Minister Hariri put Syria under renewed pressure from the international community to abide by U.N. Security Council Resolution 1559 and withdraw its troops from Lebanon. Sentence 2: Foreign ministers from all*

[7] The complete guidelines file used for training will be published upon publication.

[8] Some of the authors of the paper annotated a small set of reference gold target highlights and sentence unions.




(a) Step 1

(b) Step 2

(c) Step 3

(d) Step 4

Figure 6: The interface used for the annotation process.

*European Union (EU) member states, who gathered here for a meeting, on Wednesday urged Syria to withdraw its troops completely from Lebanon. Correct union: The February assassination of former Lebanon Prime Minister Hariri put Syria under renewed pressure from foreign ministers from all European Union (EU) member states gathered for a meeting, on Wednesday to abide by U.N. Security Council Resolution 1559 and withdraw its troops from Lebanon. Wrong union: The international community, including the European Union (EU), has put renewed pressure on Syria to abide by U.N. Security Council Resolution 1559 and withdraw its troops from Lebanon following the February assassination of former Lebanon Prime Minister Hariri.*

*The union is wrong, because it does not mention that foreign ministers gathered for a meeting on Wednesday.*

*Please generate a correct union to the following sentences:*

*Sentence 1:* <sentence 1 goes here>

*Sentence 2:* <sentence 2 goes here>

*Correct union:*

## F Training Details

We fine-tuned $T5_{large}$ and PRIMERA models for 20 epochs on a Tesla V100-SXM2-32GB GPU. We used a hyperparameter random search strategy. The learning rate was tuned within the range $[1e-8, 5e-5]$, while the batch size varied between $[8, 16, 32]$. We also explored the weight decay range of $[0, 0.5]$ and warump step range of $[0, 300]$. The best model was selected based on

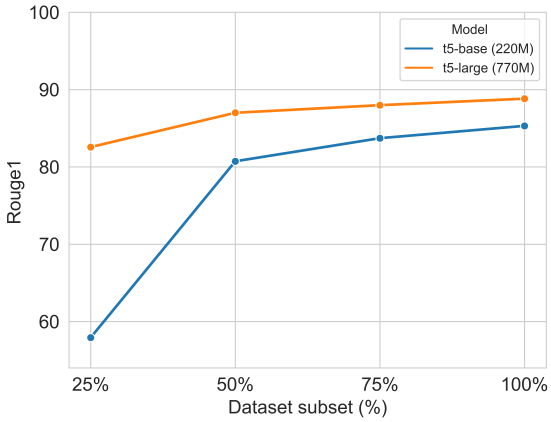



Figure 7: An evaluation of $T5$ models on different subsets of our training data $[25\%, 50\%, 75\%, 100\%]$, as well as different model sizes ($T5_{base}$ and $T5_{large}$). The number of parameters is indicated for each model.

the $ROUGE1$ metric.[9] The best T5 model was obtained with a learning rate of $4.3e-6$, no weight decay, no warmup steps, batch size of 32, after 18 epochs. For the best-performing PRIMERA model, we used a learning rate of $3.5e-6$, weight decay of 0.5, warmup steps of 80, batch size of 16 and selected the best checkpoint after 9 epochs. The training time for $T5_{large}$ and PRIMERA models were approximately 1 hour and 10 minutes each.

**Input structure** When concatenating the two source sentences to insert as input for the model, we add special separator tokens to make the model aware of the sentence boundaries. For $T5_{large}$, we separated between the source sentences in the input using a newly created special token, while for PRIMERA, we used the *<doc-sep>* token, which was used in the pre-training phase to separate between input source documents.

## G Learning Curves

To assess the adequacy of our dataset size, we evaluated the baseline models on different subsets of our training data ($[25\%, 50\%, 75\%, 100\%]$) and various model sizes ($T5_{base}$ and $T5_{large}$). Based on our findings (Figure 7), it appears that enhancing the model size from $T5_{base}$ to $T5_{large}$ results in performance improvement. However, the marginal benefit of increasing training dataset size may be limited, and further gains may not be significant.

## H Evaluation Process

As explained in Section 7, the evaluation process involves a comparative approach, whereby all the unions of system-generated sentences are evaluated simultaneously, as shown in Figure 8. The evaluation is conducted separately for four criteria. To assess the content differences between the reference union and the system union, including coverage and faithfulness, a single sentence is designated as the base sentence, and the worker is asked to evaluate the other sentence based on the amount of missing content. The reference union serves as the base sentence for evaluating coverage, while the system union is used as the base sentence for evaluating faithfulness since any information present in the system union but absent in the reference union is deemed unfaithful. In evaluating redundancy and fluency, the evaluator is only presented with the system union without the reference union.

To assess the coverage and faithfulness criteria, the workers are required to compare the generated union with the reference union, aided by red strikethroughs on words that are not included in the generated union and green highlights on words that are not included in the reference union, as illustrated in Figures 8a and 8b. For redundancy and fluency criteria, the reference union is not needed, as demonstrated in Figures 8c and 8d.

## I Example Sentence Unions

See Table 7 for examples of sentence unions, including the sentence unions from each predicted system.

## J Error Analysis

In order to perform an error analysis, we analyzed 20 examples that were rated less than perfect for all metrics based on the human evaluation (see §8.1). The findings are presented in Table 8, with one representative example from each subcategory included in Table 9. Our key observation is that models make various coverage errors as they fail to identify the uni-directional entailment correctly in the dataset. Furthermore, models make multiple coverage and faithfulness errors by incorrectly combining information and attaching it to the wrong entity or predicate.

[9]We used the HuggingFace package (Wolf et al., 2020) for both fine-tuning the models and automatically evaluating them.

### Information Coverage criterion

Any piece of information described in the reference sentence should be covered in the system sentence. Rate the following system sentences based on the amount of information they cover from the reference sentence.

**Reference sentence**

The Chernobyl, also called CIH, virus, attacks computers using Windows 95 and 98 and can overwrite their hard drives.

| # | System sentence (compared to reference) | Rank Information Coverage |
|---|---|---|
| | Use the following scale: 1 - Substantial information is missing. 2 - Some information is missing. 3 - Minor details are missing. 4 - Nothing is missing. | |
| 1 | The Chernobyl, also called CIH, virus~~, attacks~~ could erase a computer's hard drive attacked computers using Windows 95 and 98 ~~and can overwrite their~~ files, overwriting a computer's hard ~~drives~~ drive. | ○ ○ ○ ○ 1 2 3 4 |
| 2 | The Chernobyl, also called CIH, virus~~, attacks computers using Windows 95~~ could erase a computer's hard drive and ~~98 and can overwrite their hard drives~~ attacks Windows 95 and 98 files. | ○ ○ ○ ○ 1 2 3 4 |
| 3 | The Chernobyl, also called CIH, virus~~, attacks~~ could erase a computer's hard drive by attacking computers using Windows 95 and 98 ~~and can overwrite their hard drives~~ files, overwriting them. | ○ ○ ○ ○ 1 2 3 4 |

(a) Coverage

### Information Faithfulness criterion

Any piece of information described in the system sentence should be implied from the reference sentence, otherwise it is considered unfaithful. Rate the following system sentences based on the their faithfulness to the reference sentence.

**Reference sentence**

The Chernobyl, also called CIH, virus, attacks computers using Windows 95 and 98 and can overwrite their hard drives.

| # | System sentence (compared to reference) | Rank Information Faithfulness |
|---|---|---|
| | Use the following scale: 1 - Substantial information is unfaithful. 2 - Some information is unfaithful. 3 - Minor details are unfaithful. 4 - Nothing is unfaithful. | |
| 1 | The Chernobyl, also called CIH, virus~~, attacks~~ could erase a computer's hard drive attacked computers using Windows 95 and 98 ~~and can overwrite their~~ files, overwriting a computer's hard ~~drives~~ drive. | ○ ○ ○ ○ 1 2 3 4 |
| 2 | The Chernobyl, also called CIH, virus~~, attacks computers using Windows 95~~ could erase a computer's hard drive and ~~98 and can overwrite their hard drives~~ attacks Windows 95 and 98 files. | ○ ○ ○ ○ 1 2 3 4 |
| 3 | The Chernobyl, also called CIH, virus~~, attacks~~ could erase a computer's hard drive by attacking computers using Windows 95 and 98 ~~and can overwrite their hard drives~~ files, overwriting them. | ○ ○ ○ ○ 1 2 3 4 |

(b) Faithfulness

### Semantic Repetition criterion

A piece of information should be introduced only once to the reader of the sentence, unless necessary to repeat for grammaticallity reasons. Otherwise, it is considered a semantic repetition. Rate the following merged sentences based on the amount of their semantic repetition.

| # | System sentence | Rank Semantic Repetition |
|---|---|---|
| | Use the following scale: 1 - Substantial information is repeated. 2 - Some information is repeated. 3 - Minor details are repeated. 4 - Nothing is repeated. | |
| 1 | The Chernobyl, also called CIH, virus could erase a computer's hard drive attacked computers using Windows 95 and 98 files, overwriting a computer's hard drive. | ○ ○ ○ ○ 1 2 3 4 |
| 2 | The Chernobyl, also called CIH, virus could erase a computer's hard drive and attacks Windows 95 and 98 files. | ○ ○ ○ ○ 1 2 3 4 |
| 3 | The Chernobyl, also called CIH, virus could erase a computer's hard drive by attacking computers using Windows 95 and 98 files, overwriting them. | ○ ○ ○ ○ 1 2 3 4 |

(c) Repetition

### Fluency criterion

The sentence should progress fluently, form a coherent whole and it should be easy to understand the text. Rate the following merged sentences based on their fluency.

| # | System sentence | Rank Fluency |
|---|---|---|
| 1 | The Chernobyl, also called CIH, virus could erase a computer's hard drive attacked computers using Windows 95 and 98 files, overwriting a computer's hard drive. | ○ ○ ○ ○ ○ 1 2 3 4 5 |
| 2 | The Chernobyl, also called CIH, virus could erase a computer's hard drive and attacks Windows 95 and 98 files. | ○ ○ ○ ○ ○ 1 2 3 4 5 |
| 3 | The Chernobyl, also called CIH, virus could erase a computer's hard drive by attacking computers using Windows 95 and 98 files, overwriting them. | ○ ○ ○ ○ ○ 1 2 3 4 5 |

(d) Fluency

Figure 8: The interface used for the evaluation of a predicted sentence union's quality.

| | | |
|---|---|---|
| Sentence 1 | French museum officials traveled to New York last month and confirmed the find is indeed the missing Picasso work, which the Centre Georges Pompidou realized was missing from its storerooms in 2001 following a loan request; it was then valued at more than $2.5 million. | |
| Sentence 2 | The canvas had been smuggled out of a storeroom of the Centre Georges Pompidou, the Paris museum and arts center, and its whereabouts had not been known. | |
| Gold union | French museum officials traveled to New York last month and confirmed the find is indeed the missing Picasso canvas smuggled out of a storeroom of the Centre Georges Pompidou, the Paris museum and arts center, which realized it was missing in 2001 following a loan request; it was then valued at more than $2.5 million. | |
| $T5_{large}$ | French museum officials traveled to New York last month and confirmed the find is indeed the missing Picasso work, which the Centre Georges Pompidou realized was missing from its storerooms in 2001 following a loan request; it was then valued at more than $2.5 million, and its whereabouts had not been known. | Coverage: 2.0<br>Faithfulness: 4.0<br>Repetition: 3.0<br>Fluency: 3.0<br>Consolidation: 3.0 |
| PRIMERA | French museum officials traveled to New York last month and confirmed the find is indeed the missing Picasso work, which the Centre Georges Pompidou realized was missing from its storerooms in 2001 following a loan request; it was then valued at more than $2.5 million. | Coverage: 2.0<br>Faithfulness: 4.0<br>Repetition: 4.0<br>Fluency: 5.0<br>Consolidation: 3.3 |
| GPT3 | French museum officials traveled to New York last month and confirmed that the canvas, which had been smuggled out of a storeroom of the Centre Georges Pompidou in Paris and its whereabouts had not been known since 2001 following a loan request, is indeed the missing Picasso work, valued at more than $2.5 million. | Coverage: 3.0<br>Faithfulness: 4.0<br>Repetition: 4.0<br>Fluency: 2.0<br>Consolidation: 3.7 |

Table 7: Examples of predicted union sentences from each baseline system and their corresponding human evaluation.

| Subcategorization | Coverage | Faithfulness | Repetition | Subcategory Explanation |
|---|---|---|---|---|
| Uni-directional entailment | 17 | 2 | 5 | This includes cases where either the entailing part is missing and the entailed part is present in the sentence or both the entailing and entailed parts are present in the sentence. |
| Wrong attachment | 13 | 13 | 1 | This includes cases where an argument is attributed to the wrong predicate or entity. |
| Lexical similar but different information | 8 | 0 | 0 | This includes cases where information is omitted, and the omitted information had a phrase that was lexically similar to a phrase in the other sentence. |
| Ignores prefix | 4 | 0 | 0 | This includes cases where the prefix to the sentence in the source is omitted from the union. |
| Related new information | 2 | 0 | 0 | This includes cases where the source sentences contain related but different information, and one of them is not included in the union. |
| Paraphrase | 1 | 1 | 5 | This includes cases where paraphrased information from the source is repeated in the union. |
| External hallucination | 0 | 3 | 0 | This includes cases where there is information in the union that does not originate from the source sentences. |

Table 8: Error analysis based on a sample of 20 erroneous examples, each example analyzed for the 3 system outputs. For each metric, we report the frequency of a subcategory that we suspect is the cause for the error. One representative example from each subcategory is included in Table 9.

| Subcategorization | Prediction | Explanation |
|---|---|---|
| External hallucination | Peter Capaldi was revealed as the 12th Doctor of the Doctor Who series during a special live broadcast, with the announcement being made that he had been cast as the 12th Time Lord. | The mention of a live broadcast is not part of the source sentences. Interestingly, this is true, which indicates that the model knows this story. |
| Lexical similar but different information | Sgt. Tim Shields and Attorney-General Wally Oppal announced Wednesday that the RCMP arrested two Bountiful residents, Winston K. Blackmore, 52, and James Oler, 44, on charges of polygamy. | Source sentence mentioned “and leaders of a polygamist group”. This was possibly skipped due to the model incorrectly recognizing "polygamy" later as a paraphrase. |
| Uni-directional entailment | A strong 6.1-magnitude earthquake which hit the Indonesian northwestern province of Aceh on Tuesday killed a child, injured dozens and destroyed buildings, sparking panic in a region devastated by the quake-triggered tsunami of 2004. | Sentence 2 mentions “injuring at least 50 people” which entails “dozens injured” in sentence 1, but it is not mentioned in the union. |
| Ignores prefix | The 55-year-old Scottish actor Peter Capaldi is officially set to replace exiting star Matt Smith, who announced in June that he was leaving the sci-fi show later this year, as the TARDIS leader, as producer Steven Moffat announced on the live BBC special Doctor Who Live: The Next Doctor Sunday. | Ignores the information about it being the 12th doctor, which was mentioned in a sentence prefix: “Doctor Who has finally selected its 12th doctor: Peter Capaldi is officially set to ...”. |
| Related new information | Industry analysts contacted by eWEEK generally say they believe that Hewlett-Packard’s $13.9 billion acquisition of Electronic Data Systems, which was officially announced on May 13 and is currently being negotiated, is a good move for both companies, although there will be the usual integration snafus such as vendor neutrality issues, business lines, culture shock and layoffs. | "good move for both companies" and "a deal that could help the world’s largest personal computer maker snap up more data management and consulting contracts" are different, and both should be mentioned in the union. |
| Paraphrase | In France, art dealers are obliged by law to register all purchased art, except those bought at public auction. | Sentence 1 mentions “art dealers ... purchases”, and sentence 2 mentions “dealers ... purchased art”. Since these are paraphrases, the union which repeates both “art dealers” and “purchased art” is repetitive. |
| Wrong attachment | The flight recorder was recovered on November 9 and revealed that the autopilot was disconnected, the descent appeared "controlled," the cockpit turned off both engines, and the elevators were out of unison, something experienced pilots would not do. | “something experienced pilots would not do” refers to turning out both engines, not elevators out of unison. This is usually caused by an incorrect merge of the sentences. |

Table 9: Examples for the subcategories we devised during the model error analysis, which we suspect are are the cause for the error.