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] OneAI
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 and multi-hop long-form question answering, 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 Introduction

Learning a new topic or finding answers to complex questions usually requires reading multiple sources of textual information. While the information coming from a single document tends to be coherent, documents from different sources while using different lexical phrasing, often at varying levels of specificity, to convey similar information, as exemplified in Fig. 1. We refer to text consolidation as the process of taking multiple partly overlapping textual sources and merging them into a single coherent and complete form.

Many downstream tasks require multi-text consolidation, such as Multi-Document Summarization (MDS) (Narayan et al., 2018; Fabbri et al., 2019) and multi-hop long-form question answering (Fan et al., 2019). Aiming at a more controlled environment for researching such settings, a sentence fusion task was introduced in which a set of sentences is fused into a single sentence (Barzilay and McKeown, 2005; Thadani and McKeown, 2013; Weiss et al., 2021).

[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.

However, being similar to summarization, the general sentence fusion task is ill-defined, because it allows for *subjective* salience-based content selection decisions. 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 more attractive due to its more *objective* and semantically challenging nature, we found that there are relatively few datasets for it (McKeown et al., 2010; Geva et al., 2019; Lebanoff et al., 2020), none of them sufficiently addressing the text consolidation setting.

Our work therefore revisits the sentence union generation task and proposes using it as a 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 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, [1] 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 state of the art neural generative models to address the union task, assessing their current 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 (Thadani and McKeown, 2013; Du et al., 2022; 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. 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 following works mostly focused on the intersection fusion part of the dataset. Further, a dataset of 300 examples is currently not sufficient 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, where a natural language text is generated from a knowledge graph. 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 Slobodkin et al. (2022) 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 sen-

[1] The dataset is attached to the paper submission and will be published together with the code upon publication, under the CC BY-NC 4.0 (non-commercial) license.

[S1] *A report in Vanity Fair identifying 1972-73 Watergate leaker "Deep Throat" as the FBI Deputy Director W. Mark Felt, was verified by Felt's grandson on May 31, 2005.*
[S2] *Felt's grandson, Nick Jones, made the claim in a statement read to reporters outside the family home in Santa Rosa, California, following an article in Vanity Fair in which Felt told the magazine he was "Deep Throat."*

Relation Type	S1	S2
S2 entails S1 <=	*Felt's grandson*	***Felt's grandson, Nick Jones***
S1 entails S2 =>	***1972-73 Watergate leaker "Deep Throat"***	*"Deep Throat"*
	FBI Deputy Director W. Mark Felt	*Felt*
S1 equivalent to S2 <=>	***A report in Vanity Fair … was verified by***	***Made the claim … following an article in Vanity Fair***
Disjoint	***on May 31, 2005***	***in which** Felt **told the magazine***
		in a statement read to reporters outside the family home in Santa Rosa, California

Generation

[Union] *FBI Deputy Director W. Mark Felt's grandson, Nick Jones, verified the claim on May 31, 2005 in a statement read to reporters outside the family home in Santa Rosa, California, following an article in Vanity Fair in which Felt told the magazine he was 1972-73 Watergate leaker "Deep Throat."*

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 map these relations (possibly implicitly), and then include all new or more specific information (denoted by colors). Note that the word *"identifying"* from sentence 1 does not show in the table, since it is implied from sentence 2, not adding new information.

tences (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.

Given the examples in Fig. 2 we enumerate all the semantic links that are relevant for producing a union. Specifically, we observe 3 types of relations between information units in the source sentences which affects 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 (bi-directional entailment), needs to be identified and included exactly once in the union, otherwise it would be considered redundant. For example, in Figure 2, the text spans *"A report in Vanity Fair"* and *"was verified by"* are equivalent, though lexically different, to the text spans *"made the claim"* and *"following an article in Vanity Fair"*. Uni-directional entailing content pertains to aligned text spans where one text span can be implied from the other. A correct union should only include the more specific (entailing) text span, while including both the more and less specific mentions would be redundant, and including only the less specific mention would miss some information. An example in Figure 2 would be *"FBI Deputy Director W. Mark Felt"* in sentence 1 which is more specific than the text span *"Felt"* in sentence 2. Lastly, disjoint content is content mentioned in one sentence and not in the other, and must be included in the union. This case challenges the model to properly integrate disjoint information, both at the semantic and discourse levels. Sentence 1 and 2 from Figure 2 exclusively mention the time and location of the event, each providing distinct information.

For comparison, it is interesting to see that the union task provides a more comprehensive setup than the *intersection* task[2] for information consolidation. 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.

[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 $i \models i^*$.

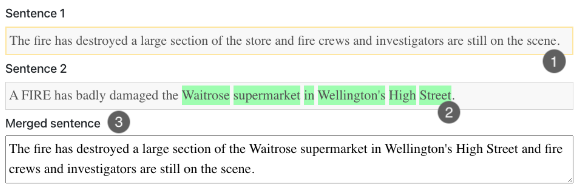


Figure 3: A screenshot of the sentence union text generation annotation interface. The screenshot shows the last step, where the worker already choose sentence 1 as the base sentence [1], highlighted the new or more specific information in sentence 2 [2] and wrote the final sentence union ("Merged sentence")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

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. We do not consider sentences with no content overlaps relevant for our dataset.

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.

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

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

Each example was given to a single annotator as we aimed to maximize the number of different inputs in our dataset, given our annotation budget. 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 Generating a coherent sentence union is sometimes unreasonable, such as when the source sentences are in disagreement about the details of an event, or context is missing in order to consolidate two overlapping pieces of information. For such cases, workers had the option to skip examples (see App. A for the guidelines).

Edge cases There are multiple edge cases concerning the source sentences that will affect the resulting sentence union. Such edge cases include world knowledge, temporal issues, subjectivity and attribution. For examples and guidelines provided to the workers for these edge cases, refer to App. B.

4.3 Cleaning annotations

To ensure a high quality dataset we introduced a post-processing step where 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 edge

Split	Train	Dev	Test	Skipped
Size	1077	347	482	465

Table 1: Sizes of the splits of our dataset, as well as of the skipped examples (19.6% 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.

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.

5.1 Sentence union quality

To estimate the reliability of our dataset, the authors of the paper have conducted a human assessment on a sample of 100 examples of sentence unions generated by our annotators. Our aim 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*). We also, additionally, 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 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 anal-

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

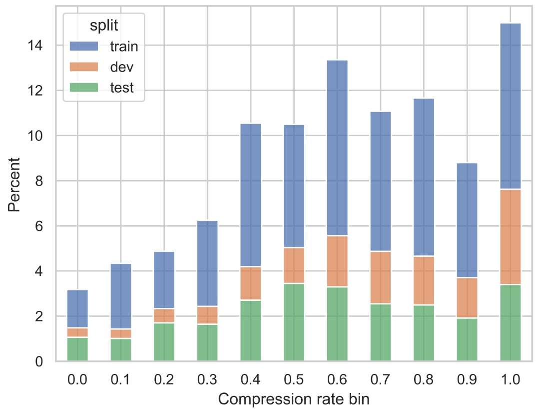


Figure 4: Compression Rate (CR) vs. the frequency of each CR bin, for the train/dev/test dataet splits.

ysis, 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.

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 60.95% (± 29.05%) for our dataset and an average CR score of 66.13% (± 23.32%) for McKeown et al. (2010). These results imply that our dataset on average contains somewhat less 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),

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

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.

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.

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. G 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* 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 ΔCR (§5.2), calculated as the average relative difference between the CRs of the predicted vs. the reference union sentences, on each instance.

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. H. 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

	Coverage (1 to 4)	Faithfulness (1 to 4)	Redundancy (1 to 4)	Consolidation (1 to 4)	Fluency (1 to 5)	$ROUGE1$	NLI	ΔCR
PRIMERA	3.3 (± 0.8)	3.6 (± 0.7)	3.8 (± 0.5)	**3.6 (± 0.4)**	4.0 (± 1.0)	**88.0% (± 8.2%)**	87.3% (± 33.3%)	**23.2% (± 21.9%)**
$GPT3$	**3.4 (± 0.7)**	3.4 (± 0.8)	3.7 (± 0.6)	3.5 (± 0.4)	3.9 (± 1.0)	85.3% (± 9.3%)	**95.6% (± 20.4%)**	25.7% (± 26.7%)
$T5_{large}$	3.0 (± 0.9)	**3.7 (± 0.6)**	**3.9 (± 0.4)**	3.5 (± 0.4)	**4.2 (± 0.9)**	87.3% (± 9.7%)	75.5% (± 43.0%)	28.6% (± 28.2%)

Table 4: Human and automatic evaluation results of system generated unions over the complete test set. All scores are averages, along with their standard deviation.

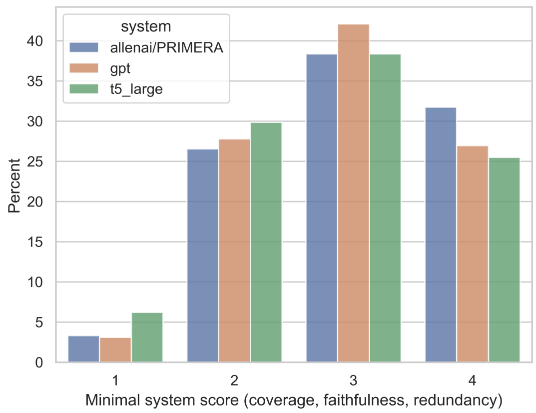


Figure 5: A histogram of minimal system scores, testing for coverage, faithfulness or redundancy mistakes.

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, and (2) the longer sentence. Based on evaluation of 50 examples completed by the authors, we report an average redundancy score of 1.6 (± 0.9) for the concatenation and an average coverage score of 2.2 (± 1.0) for the longer sentence. All the models do substantially better, with scores closer to 4, than these naive models.

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 31.7% of the predictions are fully correct with respect to content and redundancy, another 38.3% examples include minor errors, and 26.5% 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 $ROUGE1$ and the ΔCR scores are the highest for PRIMERA.

To assess our metrics quality, we follow the standard practice (Fabbri et al., 2021) and calculate a Kendall τ coefficient (McLeod, 2005) between the human and automatic evaluation results. Our results show that ΔCR and $ROUGE1$ are the highest correlated metrics with the consolidation measure ($\tau \geq 0.39$, $p < 0.05$), while $RougeL$ correlates the most with fluency ($\tau = 0.20$, $p < 0.05$). Overall, these metrics can be used jointly to provide automatic feedback when developing models.

8.3 Error analysis

To shed light on the different mistakes produced by the baseline models, we analyzed 20 erroneous examples detected in the human evaluation, with each example including 3 predictions, one per baseline system. We found that the most common causes for model mistakes are due to the richness of informational relations in the source sentences, most commonly: (1) uni-directional entailment, (2) lexically similar pieces of information that actually provide different information, and (3) coreference resolution errors, where attaching a piece of information from the other sentence can result in hallucinations. This analysis is reported in App. I.

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

As follows we enumerate some limitations to our work. While we did create the largest union dataset to date, it could be that the training data size might be still too small to fine-tune a pretrained language model. Fusion data, on the other hand, is easier to generate automatically.

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

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.

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

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 and Kathleen R. McKeown. 2005. Sentence fusion for multidocument news summarization. *Comput. Linguist.*, 31(3):297–328.

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).

Yupei Du, Qi Zheng, Yuanbin Wu, Man Lan, Yan Yang, and Meirong Ma. 2022. Understanding gender bias in knowledge base embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1381–1395, Dublin, Ireland. 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.

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.

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

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.

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.

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. 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.

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

The workers were instructed to skip examples in the following cases.

Unrelated sentences When two input sentences have little or no overlap and it is impossible to generate a coherent and comprehensible merged sentence. For example, sentence 1 is *"South Korea, the United States, Japan, China and Russia are trying to persuade North Korea to abandon its nuclear weapons development in talks that may resume as early as next week in the Chinese capital of Beijing after going into recess last month."* and sentence 2 is *"The United States and North Korea hold the first round of high-level talks in Pyongyang over North Korea's suspected construction of an underground nuclear facility."*.

Disagreements Sometimes, there are two information statements that are opposed to 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, and sentence 2 mentions that the child is 12 years old.

No text consolidation If two sentences are related, but there is no information to consolidate and their sentence union is simply a concatenation of the two sentences. For example, sentence 1 is *"Acupuncture is the ancient Chinese medical therapy technique of inserting thin, sharpened needles into specific nerve junction points of the body."* and sentence 2 is *"In the Yale study, 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."*.

Unnatural unions When unifying two input sentences will make an unnatural sentence union. For example, sentence 1 is *"Fannie Mae's board met Sunday night to discuss Raines' future."* and sentence 2 is *"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."*. Sentence 1 uses the past tense, and sentence 2 uses the future tense. It would make sense to use the past tense, because at the time of writing the sentence union the event is in the past. However, keeping the piece of information that someone said something on Friday before the event, will make the sentence union very unnatural.

B Edge Cases

This section addresses different semantic aspects that we should notice when comparing two source sentences, that will eventually affect how the sentence union is written.

Date of publication Some sentences might mention a specific time relative to the day the sentence was written. The mention *"yesterday"* is an obvious example, but also *"Monday"*, which implies that the sentence was written in the same week of the event, and *"earlier this month"*, which implies that the sentence was written in the same month of the event. The workers were instructed to assume the date of publication is known, so there is no difference between the mention of *"yesterday"* and the mention of *"Monday"*. However, the mention of *"yesterday"* is more specific than the mention of *"earlier this month"*.

Before and after an event Some sentences might differ in their time of publication compared to the event they refer to. 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."*. Workers were instructed to use the past tense, as the sentence union is written after the event.

World knowledge Some sentences might mention the the same piece of information in different levels of specificity, which requires world knowledge to identify. For example, sentence 1 mentions *"Paris"* and sentence 2 mentions *"France"*. By knowing that Paris is the capital of France, it is redundant to put the mention of France in the sentence union. However, Paris is also a city in Texas, and if sentence 1 mentions *"Paris"* and sentence 2 mentions *"Texas"*, then it is likely not redundant to put Texas inside the sentence union. This is part of the ambiguity of the task, discussed in §5. Workers were instructed to assume common world knowledge when creating the sentence union.

Attribution A subtle issue of creating a sentence union is when the source sentences makes attributions to a specific source (i.e., *"news agency reported"*). For example, sentence 1 is *"Video of Brooklyn Mother Zurana Horton being shot and killed was revealed Thursday, according to the N.Y. Daily News."* and sentence 2 is *"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."*. In this case, the new information coming from sentence 2 is attributed to the video content, which is overlapping in both sentences, and not to N.Y. Daily news, so the following sentence union is valid: *"A shocking video of Brooklyn Mother Zurana being shot and killed as she picked her daughter up from school was revealed Thursday, according to the N.Y. Daily News."*.

Another example would be a sentence that contains quotes, since changing a quote to contain more information creates an unfaithful sentence union.

C Annotation Process

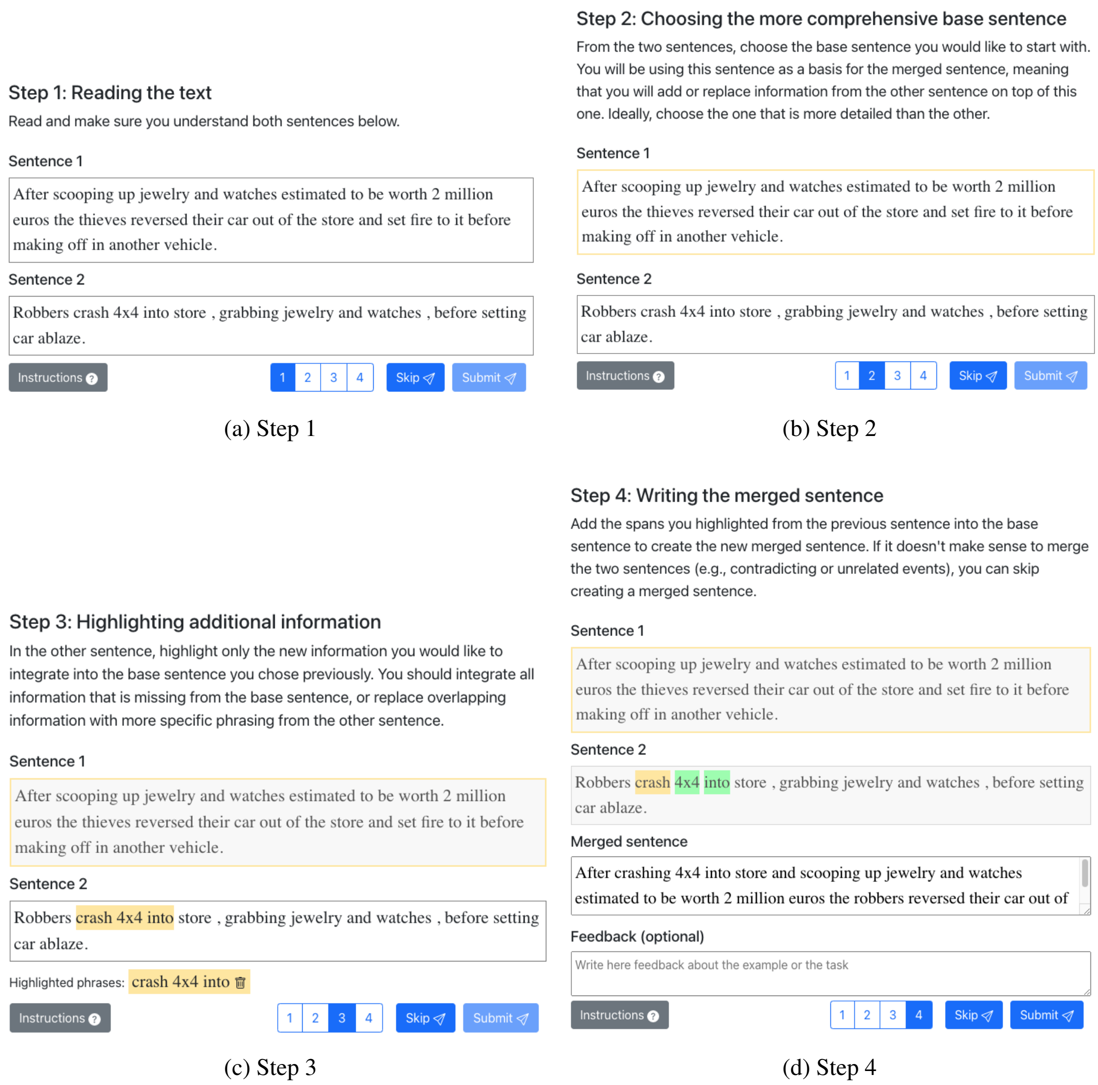
Screenshots of the entire annotation process are depicted in Figure 6. Guidelines for creating sentence unions[7] include writing one coherent sentence, ordering the information in a stand-alone manner (as if the sentence would have been written from scratch), meaning that the writing process should 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 diverse edge cases. After each annotation, workers were shown gold target highlights and sentence unions[8] for comparison with their own output.

[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.

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 assassi-

nation 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 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 $[2e-5, 4e-5]$, while the gradient accumulation steps were varied between $[2, 4, 8]$ with a fixed batch size of 8. We also explored the weight decay range of $[0, 0.5]$ with a step of 0.1. The best model was selected based on the $ROUGE1$ metric.[9] In addition, we computed the NLI score (bidirectional entailment, as described in §7) after each epoch and selected the best checkpoint according to this metric. Due to computational limitations, we did not perform the hyperparameter search based on the NLI score. The best T5 model was obtained with a learning rate of 3.1e-5, weight decay of 0.5, batch size of 8, and gradient accumulation step of 8, after 15 epochs. For the best-performing PRIMERA model, we used a learning rate of $4.2e-5$, weight decay of 0.3, batch size of 8, and gradient accumulation step of 4 and selected the best checkpoint after 11 epochs. The training time for $T5_{large}$ and PRIMERA models were approximately 1 hour and 50 minutes and 1 hour and 40 minutes respectively.

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.

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

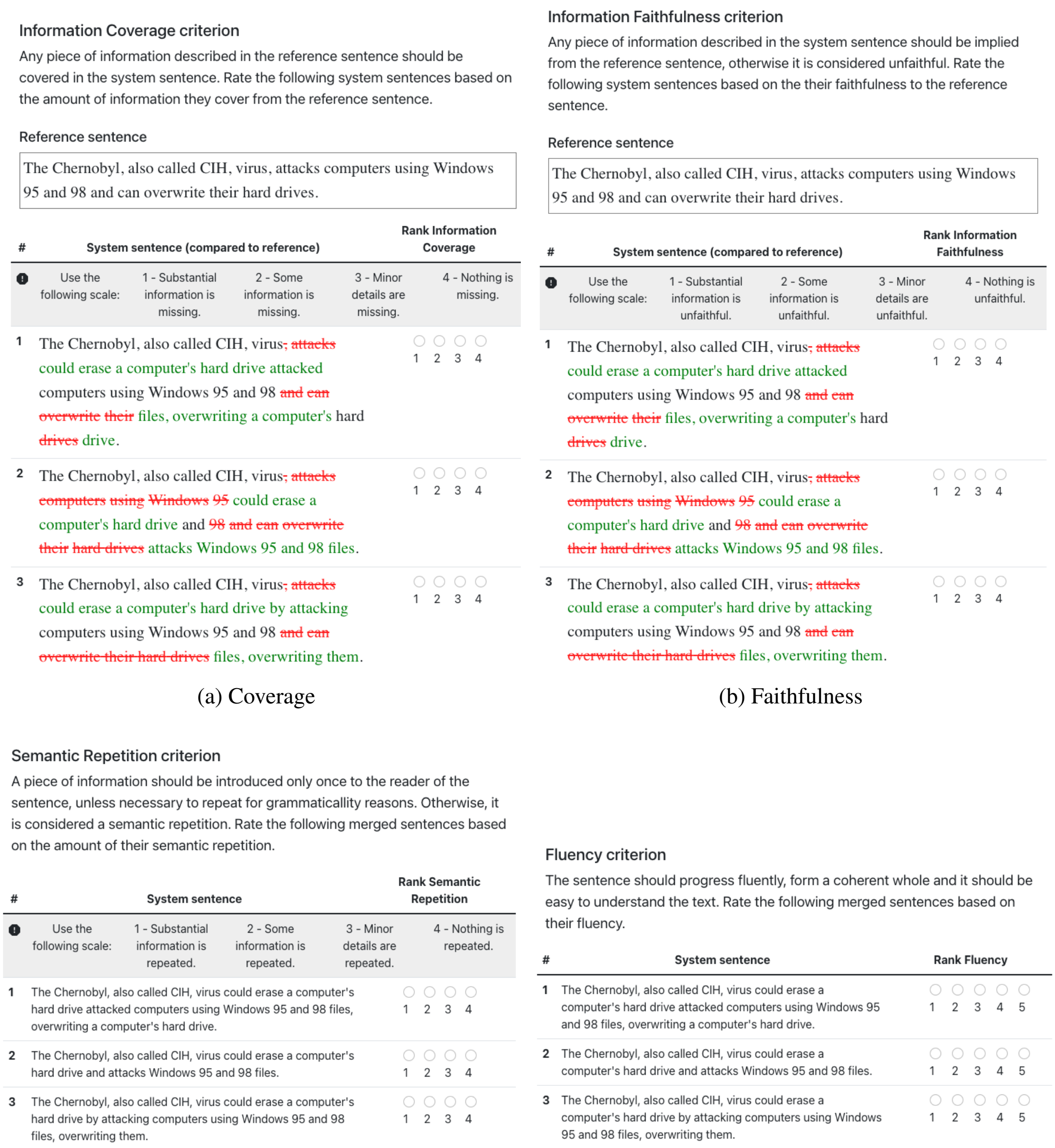
G Evaluation Process

As mentioned in §7, the evaluation works in a comparative fashion where all the system generated sentence unions are rated simultaneously (see Figure 7). This evaluation is repeated for each of the four criteria. For evaluating the content differences between the reference union and the system union (i.e., the coverage and faithfulness), we set one sentence as the *base sentence* and then ask the worker to evaluate the other sentence based on the amount of missing content. For evaluating coverage, the base sentence is the reference union, while for evaluating faithfulness, the base sentence is the system union, since information in the system union that is missing from the reference is considered unfaithful. For evaluating redundancy and fluency, the evaluator sees only the system union without the reference.

For the coverage and faithfulness criteria, the worker had to compare the generated union with the reference union. To facilitate this process, words that are not included in the generated union are marked in red with a strike-through, and words that are not included in the reference union are marked in green (see Figures 7a and 7b). For the Redundancy and Fluency criteria, the worker does not need to see the reference union (see Figures 7c and 7d).

H Example Sentence Unions

See Table 5 for examples of sentence unions, including the sentence unions from each predicted system.



(a) Coverage (b) Faithfulness

(c) Repetition (d) Fluency

Figure 7: The interface used for the evaluation of a predicted sentence union's quality.

I Error Analysis

We analyzed 20 erroneous examples based on the human evaluation in §8.1, reported in Table 6. An erroneous example is an example that received less than a perfect rating for all metrics.

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: 3.0 Faithfulness: 3.0 Repetition: 4.0 Fluency: 4.0 Consolidation: 3.3
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: 3.0 Faithfulness: 3.0 Repetition: 4.0 Fluency: 5.0 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 Faithfulness: 3.0 Repetition: 4.0 Fluency: 4.0 Consolidation: 3.3

Table 5: Examples of predicted union sentences from each baseline system and their corresponding human evaluation.

Subcategorization	Coverage	Faithfulness	Fluency	Repetition
Uni-directional entailment	19	9	12	6
Lexical similar but different information	8	0	4	2
Ambiguity	4	4	0	0
Ignores prefix	4	0	2	1
Common sense	3	3	1	0
Discourse consolidation	2	1	4	2
Internal hallucination	2	3	6	0
Related new information	2	2	2	2
Coreference	1	2	2	2
Hallucination	1	1	2	1
Lexical repetition	1	0	1	1

Table 6: Error analysis based on a sample of 20 erroneous examples, each example includes 3 systems. For each metric, we report the frequency of a subcategory that we suspect is the cause for the error.

Subcategorization	Prediction	Explanation
Discourse consolidation	Jeffs is charged with two counts of sexual assault for raping two under-aged girls and fathering a child with one of the girls, a 15 year old, after they presented a DNA report that showed he had fathered the child with the girl.	"charged ... after presenting DNA report" is incorrect, the source just mentioned "They presented a DNA report".
Hallucination	Jeffs is charged with two counts of sexual assault for raping two under-aged girls, one of which he fathered a child with according to a DNA report presented in court.	"presented in court" was not mentioned in the source.
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.
Coreference	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, as two leaders of the polygamist group were arrested in Canada.	The two leaders and the two people mentioned are the same, but the union implies they are different people.
Ambiguity	A strong 6.1-magnitude earthquake which hit the Indonesian province of Aceh on Tuesday killed at least one person, injured dozens and destroyed buildings, sparking panic in a region devastated by the quake-triggered tsunami of 2004.	One source sentence mentioned hundreds injured and the other dozens injured.
Related new information	At least 50 reported injured and buildings destroyed after a powerful 6.1 magnitude earthquake strikes Sumatra's northern Indonesian province of Aceh, killing at least three people and injuring dozens more across the affected region.	A coherent union should describe the aftermath of the quake in a sequential manner, but the predicted union separates the information with a description of the quake, probably due to failed alignment.
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 ...".
Lexical repetition	Show boss Steven Moffat announced on the live BBC special Doctor Who Live: The Next Doctor Sunday that Peter Capaldi was the only name considered to take over Doctor Who role from Matt Smith and is officially set to replace exiting star Matt Smith as the TARDIS leader.	Repeates "Matt Smith".
Common sense	When Matt Smith, 26, who played a young political researcher in the show, stepped into Doctor Who's title role as the 11th Doctor this weekend, he went from British character actor to potential entertainment legend.	It is likely that Matt Smith played in a different show, and not in Doctor Who.
Internal hallucination	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.

Table 7: Examples for the subcategories we devised during the error analysis, which we suspect are are the cause for the error.