Extending Multi-Text Sentence Fusion Resources via Pyramid Annotations

Daniela Brook Weiss[1] **Paul Roit**[1] **Ori Ernst**[1] **Ido Dagan**[1]
[1]Computer Science Department, Bar-Ilan University
{dani.b.weiss,plroit,oriern}@gmail.com dagan@cs.biu.ac.il

Abstract

NLP models that process multiple texts often struggle in recognizing corresponding and salient information that is often differently phrased, and consolidating the redundancies across texts. To facilitate research of such challenges, the *sentence fusion* task was proposed, yet previous datasets for this task were very limited in their size and scope. In this paper, we revisit and substantially extend previous dataset creation efforts. With careful modifications, relabeling and employing complementing data sources, we were able to more than triple the size of a notable earlier dataset. Moreover, we show that our extended version uses more representative texts for multi-document tasks and provides a more diverse training-set, which substantially improves model performance.

1 Introduction

Despite recent advances reported in NLU benchmarks for single document tasks, cross-document tasks, such as multi-document summarization (MDS) have not progressed with the same pace. The handling of information across documents requires effective measures for identifying overlapping content. Moreover, generative tasks require consolidating the relevant and redundant content into a coherent utterance. In light of this, several works proposed a focused sentence-level task, called *sentence fusion*, which focuses on summarizing multiple sentences with overlapping content into a non-redundant one. This allows a fine-grained analysis of which information units are shared among the input sentences, as well as control over different degrees of information inclusion and exclusion.

However, the available datasets for fusing sentences which exhibit significant content overlap are still lacking, with the most recent containing only several hundreds of examples (McKeown et al., 2010; Thadani and McKeown, 2013), impeding further research. In this work, we follow Thadani and McKeown (2013) and extend their described sentence fusion dataset, which is derived from expertly written and annotated summaries based on the Pyramid MDS evaluation method by Nenkova and Passonneau (2004). Table 1 illustrates an example where the gold label is a summary of the content intersection in the input sentences.

a.	**Fisheries in parts of the Philippines have been decimated by the use of cyanide** in fishing.
b.	Philippine fishermen use **cyanide in fishing**, needlessly **destroying immature fish**.
c.	Sodium **cyanide use by fisherman decimates fish**.
d.	In the Philippines some fishermen use homemade explosives and **cyanide for driving fish away from reefs and into nets**.
Label	Sodium cyanide use by fisherman decimates fish

Table 1: Sentence fusion example from Thadani and McKeown (2013). (a-d) are the input sentences, originating from different documents. Text spans (in bold) that are considered as contributing to the same unit of content (SCU) are annotated with the same concise label. The sentences where the spans appear in are grouped to be input for sentence fusion, while the SCU label becomes the fusion target.

We find that the heuristics and filters applied by Thadani and McKeown (2013) result in short and highly related sentences, which may not reflect more complex and long sentences that are often found in multi-text consolidation tasks. Moreover, their dataset uses exclusively sentences from expert summaries and exclude the actual source documents that are used in practice for summarization. The resulting high similarity within input sentences makes them amenable to extractive methods, where a representative sentence can be selected as the summary, curbing the efforts to develop an abstractive fusion of sentences.

In this work, we modify Thadani and McKeown (2013)'s pre-processing pipeline after careful anal-

ysis, re-label a portion of the instances, and supplement the data with source document sentences (§3). Our contribution therefore is an extended sentence-fusion dataset[1], more than 4x times larger than its original, with 18% manually relabeled instances. We show that our final extended dataset better reflects challenges in multi-source summarization tasks (§4), with highly redundant salient content, originating in more representative sentences from the wild. In addition, we show (§5) that a contemporary generative model produces more abstractive output after training on our extended training set than on the original one. Similarly, it also outperforms the latter on the original test set. Given that sentence fusion was originally motivated as a step in modular multi-document summarization pipelines (Barzilay and McKeown, 2005; Marsi and Krahmer, 2005), we hope that progress on sentence fusion may contribute to broader contexts of multi-document consolidation and fusion tasks.

2 Background

The sentence fusion task deals with combining multiple sentences with overlapping content into a single summary sentence that represents the shared information across the inputs (Barzilay and McKeown, 2005; Filippova and Strube, 2008; Marsi and Krahmer, 2005; McKeown et al., 2010; Thadani and McKeown, 2013). Several other variants of sentence fusion have also been explored, such as sentence union – fusing the union of information in the input (Marsi and Krahmer, 2005). In another variant, "disparate" sentence fusion (Elsner and Santhanam, 2011; Geva et al., 2019; Lebanoff et al., 2019, 2020), the input sentences do not exhibit considerable content overlap but are rather related in discourse. Such sentences often originate in a single document and pose a different kind of challenge to generate the right discourse structure that will fluently connect the inputs.

For pragmatic purposes, a "loose" variant of sentence intersection is desired, since redundant content is most likely salient, yet additional important but non-overlapping information may be relevant for a final summarized sentence. For this reason, our extended dataset follows the fusion as "loose" intersection approach applied by Barzilay and McKeown (2005), McKeown et al. (2010) and Thadani and McKeown (2013). The latter compiled a dataset for sentence fusion by leveraging annotations made during post-hoc evaluation of multi-document summarization systems.

A	Statoil's internal investigation acknowledged inadequate planning and a lack of risk appreciation led to the leak.
B	Statoil admitted the leak resulted from inadequate planning and appreciation of risk, and failure to observe governing documentation.
Label	Statoil admitted responsibility for the leak

Table 2: Originally filtered SCU instance in PYRFUS. This example was excluded due to the SCU contributing spans in A and B being much longer than the label itself.

2.1 From MDS Evaluation to Fusion

The Pyramid method (Nenkova and Passonneau, 2004), is a well-known evaluation method for content selection in summarization, which was used in the DUC[2] and TAC[3] benchmarks for MDS.

Applying this method, a set of reference summaries per topic are written by expert annotators and divided into informational units. Each unit, named Summary Content Unit (*SCU*), denotes a short statement. For example, the SCU labeled: *cyanide use by fisherman decimates fish* may be expressed in multiple summaries and source documents under different manifestations. To compile a list of content units for MDS evaluation, the annotator marks text spans (see bold spans in Table 1) across reference summaries with equivalent content that directly expresses or contributes to the summary unit (*SCU contributors*). Next, she labels the content unit by writing a concise statement in natural language, named *SCU Label*. The source sentences of each contributing span may then be grouped into a cluster bearing the same SCU label. Table 1 presents an example of such a cluster, containing four source sentences with contributing spans (in bold), along with their associated SCU label that concisely summarizes them. Thadani and McKeown (2013) creates a fusion instance by using each cluster's sentences as input, and the SCU label as the target for fusion output.

[1]Our Code and data can be found here: https://github.com/DanielaBWeiss/Extending-Sentence-Fusion-Resources

[2]https://www-nlpir.nist.gov/projects/duc/data.html, years used 2005-2008

[3]https://tac.nist.gov/, years used 2009-2011

3 Data Collection

After carefully analyzing Thadani and McKeown (2013)'s pre-processing pipeline described in subsection 3.1, we decided to substantially modify it (subsection 3.2), recovering significantly more data. We proceed with relabeling some of the targets, and adding samples from source documents (i.e. not just from expertly written summaries) that were mapped to SCUs but overlooked in the past.

3.1 Previous Pre-processing pipeline

Thadani and McKeown (2013) applied several filtering steps to generate a fusion dataset (termed here PYRFUS) from SCUs. Specific details regarding these filters as well as examples are in Appendix A. While the original intention was to reduce noisy samples, these steps removed a significant portion of challenging fusion instances. Potential clusters were removed for having differences in length either between the source sentences to the marked span contributions, or between the target SCU label and the marked span contributions, denoting possibly non-shared information that appeared in the input, but not in the output. Another filtering criterion had been to discard all clusters whose target label contains content words unused by any input sentence, discouraging paraphrasing between input and output. Such filtering has left the dataset, whose inputs and outputs are quite similar in both length and content (see §4), missing realistic challenges in a multi-document setup, where lexically differing and non-overlapping content may appear. Moreover, such setup inadvertently biases generative models to be more extractive (see §5) than abstractive, relying on a single input sentence to convey all shared information in a cluster.

This dataset was the largest available source to date for supervised sentence fusion focused on multi-text, with a total of 1705 fusion instances.[4]

3.2 Extending Fusion Dataset

We discovered that most of the above filtering criteria were safe to forgo save a few sanity checks. This has recovered new fusion instances by either adding back removed SCU labels or input sentences. Following, we noticed that 18% of the input clusters share more than one SCU label, mostly due to the original Pyramid annotators splitting conjunctions along different SCUs. For correctness, we manually re-labeled such clusters using all shared labels into a single sentence (see Appendix C).

Additionally, DUC also made available the SCU Marked Corpus (Copeck and Szpakowicz, 2005), which automatically maps *source* document sentences to SCU labels using lexical matching. We use this resource to extend our dataset with document sentences, which were overlooked in PYRFUS. Document sentences tend to be longer and more varying than summary sentences, with 30 tokens vs. 20 on average. Clusters containing document sentences also tend to have more inputs, since reference summaries were limited to four, while the number of source documents per summarized topic is much higher. In total, we have extended the fusion dataset from its original 1705 instances to 7505, with 37% containing at least one document source sentence, creating a much more varied dataset, as analyzed next.

Fusion Data	Total	Avg Clus.	R_1 L-to-S	R_1 S-to-S
DISPARATE	**1599**	2	32.7	15.0
PYRFUS	**1705**	2.8	46.5	35.0
Δ-PYRFUS	**5842**	3.3	34.6	31.6
PYRFUS++	**7505**	3.3	37.8	32.2

Table 3: Comparisons of fusion datasets and variations. DISPARATE (Lebanoff et al., 2020) introduced a disparate-fusion dataset, containing exactly 2 input sentences. L-to-S and S-to-S refer to label-to-sentence and sentence-to-sentence ROUGE scores, respectively.

4 Data Analysis

We suggest that the additional instances previously skipped would more closely resemble challenges in a multi-document setting. To show that, we compare our extended dataset PYRFUS++ to its predecessor PYRFUS, that uses closely knit sentence clusters, and to DISPARATE (Lebanoff et al., 2020), that contains mostly non-overlapping within *document* sentences. The latter allows to estimate a lower bound for overlap for document sentences with little shared content that still relate to each other, making the bound tighter than for randomly picked sentences (some examples are shown in Appendix D). We denote by Δ-PYRFUS the instances that we added exclusively as part of our extension.[5] To assess content overlap empirically, we calculate the micro-average of ROUGE (Lin, 2004) word-

[4] This count was reproduced using the author's published code, the originally reported count is slightly higher.

[5] The original clusters may have grown as well due to the added input sentences, but we exclude those clusters from Δ-PYRFUS for ease of analysis

Train Data	Dev	Test	Test++
PYRFUS	36.4	40.9	28.5
PYRFUS++	**42**	**45.4**	**32.5**

Table 4: Rouge-2 F1 results for the baseline model (BART). Test++ refers to the test set of the extended PYRFUS++ dataset. The other evaluation splits refer to the original PYRFUS data.

SCU Label	Sodium cyanide use by fisherman **decimates** fish
PYRFUS	In the Philippines some fishermen use cyanide in fishing
PYRFUS++	In the Philippines cyanide use by fisherman **decimates** fish

Table 5: The gold SCU label vs the predictions made by the baseline model trained on PYRFUS and PYRFUS++

overlap between every sentence in the cluster to its target label ($R_1^{L\to S}$) and between every pair of input sentences in the same cluster ($R_1^{S\to S}$).

The results in Table 3 show that the content overlap among input sentences ($R_1^{S\to S}$) of our added instances in Δ-PYRFUS is much closer to PYRFUS than to disparate sentences, indicating they are viable and highly-related input examples. This reinforces our claim that in a true multi-document setting a system will be challenged with dealing with significantly more redundant information then exhibited within a single document (as in DISPARATE), and this has to be specifically addressed by a multi-document fusion dataset.

As expected, PYRFUS contains a much higher label to sentence content overlap ($R_1^{L\to S}$), given that the original pre-processing explicitly removed instances with less overlap between the SCU output and the source sentences. In fact, our analysis revealed that in PyrFus, extractive target labels, where the target sentence is an approximate copy of one of the input sentences (up to two words), account for 29% of the clusters, while in our extended dataset they account only for 11%. Overall our new fusion clusters express high relatedness between the source sentences and their label, while exhibiting higher diversity.

5 Baselines and Data Effectiveness

We implement a modern baseline (see Appendix E for details) for PYRFUS (Thadani and McKeown, 2013), which outperforms their pre-neural one.[6] To that end, we employ the pre-trained auto-encoder BART (Lewis et al., 2020) as our end-to-end generation model due to its demonstrated performance on summarization tasks.

Results, as shown in Table 4, were measured with the Rouge-2 F1 metric on the original PYRFUS evaluation splits. These results show that a fusion model trained on our extended data (PYRFUS++) significantly outperforms the same model trained on the original training data, by roughly 5 R_2 points. Notably, the model trained on PYRFUS++ scored 13 points lower on its own test set, indicating that the new dataset is much more challenging, and yet enables the model to reach better generalizations.

Examining the outputs of both models, we find that many are similar and are often extracted from source sentences[7]. To study the differences between model outputs, we first sample 50 instances where the PYRFUS++ model performed worse. We notice that in most (78%) of these cases the PYRFUS++ model output is acceptable, while the lower score stems from ROUGE artifacts due to sentence rephrasing. Only 22% do suffer from lack of salient content. On the contrary, inspecting 50 instances where the PYRFUS model performed worse, we find that only 54% of these are acceptable, while the rest suffer from lack of salient content. This sample analysis suggests that the advantage of the PYRFUS++ model is even greater than reflected by the ROUGE scores. Finally, an example for missing information in the PYRFUS model output appears in Table 5, not including a critical detail that all input sentences (in Table 1) discuss – fish decimation, while the PYRFUS++-trained model correctly includes it. Such instances show the necessity of a large and realistic fusion dataset for model training.

6 Conclusion

In this work we extended a sentence fusion dataset by almost four times its original size, while relabeling some of the data. The new dataset includes more complex and relevant training instances, better reflecting those that could be found in "the wild", and thus facilitates further research on data consolidation in multi-text tasks. In addition, we

[6] PyrFus evaluation used bigram-F1 (Unno et al., 2006) that is similar to Rouge-2 F1, reporting 24.92 points for their best model. We use the widely accepted Rouge metric to be inline with contemporary works.

[7] A fairly large proportion of targets are inevitably extractive, since the original (manual) Pyramid data contains many extractive SCU labels (Thadani and McKeown, 2013)

train baseline fusion models and show that when trained on our extended data we achieve notably better performance on the original available fusion test set, while also generating qualitatively better ("loose") sentence intersections.

Acknowledgements

The work described herein was supported in part by grants from Intel Labs, Facebook, the Israel Science Foundation grant 2827/21, and by a grant from the Israel Ministry of Science and Technology.

References

Regina Barzilay and Kathleen McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31:297–328.

Terry Copeck and Stan Szpakowicz. 2005. Leveraging pyramids. In *Text Summarization Branches Out*. The Workshop on Automatic Summarization (DUC 2005).

Micha Elsner and Deepak Santhanam. 2011. Learning to fuse disparate sentences. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 54–63, Portland, Oregon. Association for Computational Linguistics.

Katja Filippova and Michael Strube. 2008. Sentence fusion via dependency graph compression. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 177–185, Honolulu, Hawaii. Association for Computational Linguistics.

M. Geva, E. Malmi, I. Szpektor, and J. Berant. 2019. DiscoFuse: A large-scale dataset for discourse-based sentence fusion. In *North American Association for Computational Linguistics (NAACL)*.

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, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Scoring sentence singletons and pairs for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Erwin Marsi and Emiel Krahmer. 2005. Explorations in sentence fusion. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

Kathleen McKeown, Sara Rosenthal, Kapil Thadani, and Coleman Moore. 2010. Time-efficient creation of an accurate sentence fusion corpus. pages 317–320.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. 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.

Yuya Unno, Takashi Ninomiya, Yusuke Miyao, and Jun'ichi Tsujii. 2006. Trimming CFG parse trees for sentence compression using machine learning approaches. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 850–857, Sydney, Australia. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. 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.

A Modifications to the preprocessing pipeline

Thadani and McKeown (2013) applied several preprocessing steps to generate a fusion dataset from

	Filter	Filtered SCU Labels
1	SL	Saudi Arabia urged withdrawal
2	NV	Resignation of Prime Minister Karami and his government
3	NV ∣ SL	Murder in Boulder, Colorado
4	NV	Confirmed bird flu cases in Hong Kong
5	SL	FARC commits slaughters
6	SL	Water being diverted

Table 6: Originally filtered SCU instances in PYRFUS. NV – No Verb, SL – Short Label. Ex. 1-4 are clusters that were excluded based on the label alone in the original dataset, but kept in our extension. Ex. 5-6 were discarded in both datasets.

SCUs. These include *discarding* all clusters that: [1] have more than 4 contributing sentences; [2] have SCU labels that don't contain a verb after the first token; [3] have SCU labels and source sentences with less than 5 words or more than 100; [4] have contributing spans that are shorter than half of their source sentence; [5] have SCU labels that are shorter than half of the shortest contributing span in the input; and [6] have SCU labels with any tokens not appearing in any of the source sentences.

Table 6 represents examples of fusion instances that were filtered out in PYRFUS due to various filters. First four examples were recovered in our dataset, but the last two were deemed too short and not specific enough, and were left out due to lack of informativeness.

We found that certain filters were safe to remove. We discard filter [2] since the majority of SCU labels without a verb use a nominalization (affecting 601 instances). Similarly, we ease the length requirement of SCUs to be between 4-100, as they were found to be coherent and descriptive, affecting 497 instances. Additionally, we allow SCU clusters that have low overlap between their label and their marked contributing spans, discarding filters [4] and [5], affecting 2410 instances (see Table 2). And finally, we keep fusion instances whose SCU label tokens are not fully covered by their input sentences to allow paraphrases (affecting 2410 instances). [8]

B Pyramid-based Fusion Data

For the fusion instances containing summary source sentences as inputs, we used the same years reported in Thadani and McKeown (2013) (years 2005-2007 for DUC and 2008-2011 for TAC). The source document sentences found in Copeck and Szpakowicz (2005) were made available from 2005-2008. We made use of all the years except 2005, since we found this year to be containing more varied documents within a topic, which yielded noisier automatic alignments between SCU labels and source document sentences.

C Manual Target Re-annotation

Once we removed most of the filtering pipeline of PYRFUS, we noticed that almost 20% of the fusion input clusters share more than one SCU label. To accommodate, we manually re-label such clusters using all shared SCU labels into a single sentence. For example, for the following two SCU labels: *Clinical trials typically involve three phases* and *Clinical trials involve an average of 200 patients per trial*, a new merged fusion label would be: *Clinical trials typically involve three phases and an average of 200 patients per trial*.

D Examples of Fusion Instances

Table 7 presents 3 examples of fusion instances originating from different datasets. As previously mentioned, DISPARATE fusion involves the fusion of input sentences that often originate from a single document, containing little content overlap but related in discourse. The data used was taken from Lebanoff et al. (2020) and included 1599 samples. PYRFUS and PYRFUS++ contain examples originating from multi-text settings, while PYRFUS contains only inputs from reference summaries, and PYRFUS++ is enhanced with document source sentences, along with more complex examples.

E Training a Fusion Baseline

As described in section 5, we train two sentence fusion baselines using a pre-trained auto-encoder BART base model (Lewis et al., 2020), on PYRFUS and PYRFUS++ respectively. We used the training script [9] made available by the transformers library (Wolf et al., 2020) with the following parameters: 4 training epochs and a learning rate of 3e-5. A "steps" evaluation parameter was used with 5000 evaluation steps and an evaluation beam of 6. Max source input was limited to 265 while max target length was set to 30. Minimum target length were

[8] Filters are not mutually exclusive, therefore there can be an instance that is affected by multiple filters.

[9] https://github.com/huggingface/transformers/blob/master/examples/legacy/seq2seq/finetune_trainer.py

Source	Fusion Inputs	Fused Output
DISPARATE	(A) **The bodies** showed signs of torture. (B) **They** were left on the side of a highway in Chilpancingo, about an hour north of the tourist resort of Acapulco in the state of Guerrero.	The bodies of the men, which showed signs of torture, were left on the side of a highway in Chilpancingo.
PYRFUS++	(A) Secret zombie networks, called botnets, infect up to millions of personal computers and **countries such as China restrict internet usage.** (B) **China and Iran censor the Internet against subversion or immorality.**	China uses censorship to fight internet crimes.
PYRFUS	(A) Overseas **hackers accessed confidential information from South Korea.** (B) **South Korea's presidential mansion came under attack during 2008** **from overseas hackers**.	Hackers accessed information from South Korea.

Table 7: Examples of different fusion instances from separate sources. PYRFUS and PYRFUS++ contain sentences originating from a multi-text setting, with related events expressed redundantly and differently, motivating information consolidation. The DISPARATE fusion dataset (Lebanoff et al., 2020) uses related sentences originating in the same document, but do not necessarily carry redundancies. Instead, it tries to model the correct discourse structure that can fuse the inputs into one sentence.

set to 4, given our minimum requirements for fusion labels. The final evaluated score reported was an average score over 20 different trained models. This is due to BART being highly sensitive to the ordering of the input sentences. Both baseline models were trained using the train/test splits that were reported in Thadani and McKeown (2013), using DUC years 2005-2007 for test, TAC 2011 for dev, and TAC 2008-2010 for train.