Analyzing Sentence Fusion in Abstractive Summarization

Logan Lebanoff[♠*] **John Muchovej**[♠*] **Franck Dernoncourt**[♣]
Doo Soon Kim[♣] **Seokhwan Kim**[♡] **Walter Chang**[♣] **Fei Liu**[♠]

[♠]University of Central Florida [♣]Adobe Research [♡]Amazon Alexa AI

{loganlebanoff, john.muchovej}@knights.ucf.edu feiliu@cs.ucf.edu
{dernonco,dkim,wachang}@adobe.com seokhwk@amazon.com

Abstract

While recent work in abstractive summarization has resulted in higher scores in automatic metrics, there is little understanding on how these systems combine information taken from multiple document sentences. In this paper, we analyze the outputs of five state-of-the-art abstractive summarizers, focusing on summary sentences that are formed by sentence fusion. We ask assessors to judge the grammaticality, faithfulness, and method of fusion for summary sentences. Our analysis reveals that system sentences are mostly grammatical, but often fail to remain faithful to the original article.

1 Introduction

Modern abstractive summarizers excel at finding and extracting salient content (See et al., 2017; Chen and Bansal, 2018; Celikyilmaz et al., 2018; Liu and Lapata, 2019). However, one of the key tenets of summarization is consolidation of information, and these systems can struggle to combine content from multiple source texts, yielding output summaries that contain poor grammar and even incorrect facts. Truthfulness of summaries is a vitally important feature in order for summarization to be widely accepted in real-world applications (Reiter, 2018; Cao et al., 2018b). In this work, we perform an extensive analysis of summary outputs generated by state-of-the-art systems, examining features such as truthfulness to the original document, grammaticality, and method of how sentences are merged together. This work presents the first in-depth human evaluation of multiple diverse summarization models.

We differentiate between two methods of shortening text: sentence compression and sentence fusion. Sentence compression reduces the length of a *single* sentence by removing words or rephrasing parts of the sentence (Cohn and Lapata, 2008; Wang et al., 2013; Li et al., 2013, 2014; Filippova et al., 2015). Sentence fusion reduces *two or more* sentences to one by taking content from each sentence and merging them together (Barzilay and McKeown, 2005; McKeown et al., 2010; Thadani and McKeown, 2013). Compression is considered an easier task because unimportant clauses within the sentence can be removed while retaining the grammaticality and truth of the sentence (McDonald, 2006). In contrast, fusion requires selection of important content and stitching of that content in a grammatical and meaningful way. We focus on sentence fusion in this work.

We examine the outputs of five abstractive summarization systems on CNN/DailyMail (Hermann et al., 2015) using human judgments. Particularly, we focus on summary sentences that involve sentence fusion, since fusion is the task that requires the most improvement. We analyze several dimensions of the outputs, including faithfulness to the original article, grammaticality, and method of fusion. We present three main findings:

- 38.3% of the system outputs introduce incorrect facts, while 21.6% are ungrammatical;
- systems often simply concatenate chunks of text when performing sentence fusion, while largely avoiding other methods of fusion like entity replacement;
- systems struggle to reliably perform complex fusion, as entity replacement and other methods result in incorrect facts 47–75% of the time.

2 Evaluation Setup

Evaluation of summarization systems relies heavily on automatic metrics. However, ROUGE (Lin, 2004) and other n-gram based metrics are limited in evaluation power and do not tell the whole story (Novikova et al., 2017). They often focus on informativeness, which misses out on important facets

[*] These authors contributed equally to this work.

Proceedings of the 2nd Workshop on New Frontiers in Summarization, pages 104–110
Hong Kong, China, November 4, 2019. ©2019 Association for Computational Linguistics

System	ROUGE			Created By				Avg Summ
	R-1	R-2	R-L	Compress	Fuse	Copy	Fail	Sent Len
PG (See et al., 2017)	39.53	17.28	36.38	63.14	6.44	**30.24**	0.18	15.7
Novel (Kryciski et al., 2018)	40.19	17.38	37.52	71.25	19.77	5.39	3.59	11.8
Fast-Abs-RL (Chen and Bansal, 2018)	40.88	17.80	**38.54**	**96.65**	0.83	2.21	0.31	15.6
Bottom-Up (Gehrmann et al., 2018)	41.22	18.68	38.34	71.15	16.35	11.76	0.74	10.7
DCA (Celikyilmaz et al., 2018)	**41.69**	**19.47**	37.92	64.11	23.96	7.07	4.86	14.5
Reference Summaries	-	-	-	60.65	**31.93**	1.36	**6.06**	**19.3**

Table 1: Comparison of state-of-the-art summarization systems. Middle column describes how summary sentences are generated. *Compress*: single sentence is shortened. *Fuse*: multiple sentences are merged. *Copy*: sentence is copied word-for-word. *Fail*: did not find matching source sentences.

of summaries such as faithfulness and grammaticality. In this paper we present a thorough investigation of several abstractive summarization systems using human evaluation on CNN/DailyMail. The task was accomplished via the crowdsourcing platform Amazon Mechanical Turk. We particularly focus on summary sentences formed by sentence fusion, as it is arguably a harder task and is a vital aspect of abstractive summarization.

2.1 Summarization Systems

We narrowed our evaluation to five state-of-the-art summarization models[1], as they represent some of the most competitive abstractive summarizers developed in recent years. The models show diversity across several dimensions, including ROUGE scores, abstractiveness, and training paradigm. We briefly describe each system, along with a comparison in Table 1.

- **PG** (See et al., 2017) The pointer-generator networks use an encoder-decoder architecture with attention and copy mechanisms that allow it to either generate a new word from the vocabulary or copy a word directly from the document. It tends strongly towards extraction and copies entire summary sentences about 30% of the time.
- **Novel** (Kryciski et al., 2018) This model uses an encoder-decoder architecture but adds a novelty metric which is optimized using reinforcement learning. It improves summary novelty by promoting the use of unseen words.
- **Fast-Abs-RL** (Chen and Bansal, 2018) Document sentences are selected using reinforcement learning and then compressed/paraphrased using an encoder-decoder model to generate summary sentences.
- **Bottom-Up** (Gehrmann et al., 2018) An external content selection model identifies which words from the document should be copied to the summary; such info is incorporated into the copy mechanism of an encoder-decoder model.
- **DCA** (Celikyilmaz et al., 2018) The source text is divided among several encoders, which are all connected to a single decoder using hierarchical attention. It achieves one of the highest ROUGE scores among state-of-the-art.

2.2 Task Design

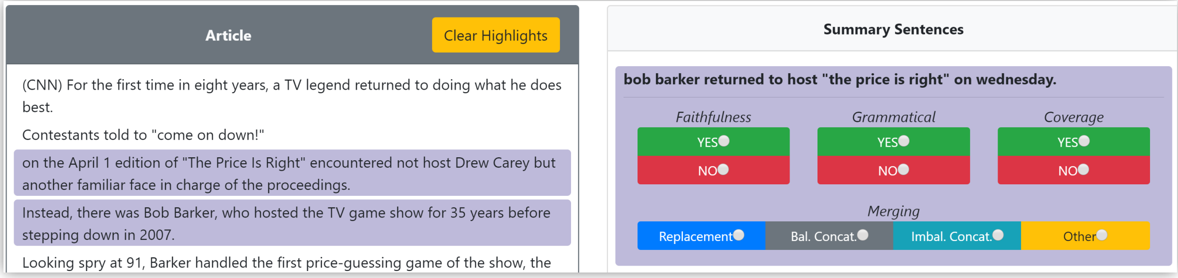
Our goal is to assess the quality of summary sentences according to their grammaticality, faithfulness and method of fusion. We design a crowd task consisting of a single article with six summary sentences: one sentence is guaranteed to be from the reference summary, the other five are taken from system summaries. An annotator is instructed to read the article, then rate the following characteristics for each summary sentence:

Faithfulness For a summary to be useful, it must remain true to the original text. This is particularly challenging for abstractive systems since they require a deep understanding of the document in order to rephrase sentences with the same meaning.

Grammaticality System summaries should follow grammatical rules in order to read well. Maintaining grammaticality can be relatively straightforward for sentence compression, as systems generally succeed at removing unnecessary clauses and interjections (See et al., 2017). However, sentence fusion requires greater understanding in order to stitch together clauses in a grammatical way.

Method of Merging Each summary sentence in our experiments is created by fusing content from two document sentences. We would like to understand how this fusion is performed. The following possibilities are given:

[1] The summary outputs from PG, Bottom-Up, and Fast-Abs-RL are obtained from their corresponding Github repos. Those from Novel and DCA are graciously provided to us by the authors. We thank the authors for sharing their work.



(CNN) For the first time in eight years, a TV legend returned to doing what he does best.

Contestants told to "come on down!"

on the April 1 edition of "The Price Is Right" encountered not host Drew Carey but another familiar face in charge of the proceedings.

Instead, there was Bob Barker, who hosted the TV game show for 35 years before stepping down in 2007.

Looking spry at 91, Barker handled the first price-guessing game of the show, the

Summary Sentences

bob barker returned to host "the price is right" on wednesday.

Figure 1: Annotation interface. A sentence from a random summarization system is shown along with four questions.

- *Replacement:* a pronoun or description of an entity in one sentence is replaced by a different description of that entity in the other sentence.
- *Balanced concatenation:* a consecutive part of one sentence is concatenated with a consecutive part of the other sentence. The parts taken from each sentence are of similar length.
- *Imbalanced concatenation:* similar to the case of "balanced concatenation," but the part taken from one sentence is larger than the part taken from the other sentence.
- *Other:* all remaining cases.

Coverage An annotator is asked to rate how well highlighted article sentences "covered" the information contained in the summary sentence. Two article sentences that best match a summary sentence are selected according to a heuristic developed by Lebanoff et al. (2019). The same heuristic is also used to determine whether a summary sentence is created by compression or fusion (more details later in this section). Given the importance of this heuristic for our task, we would like to measure its effectiveness on selecting article sentences that best match a given summary sentence.

We provide detailed instructions, including examples and explanations. We randomly select 100 articles from the CNN/DailyMail test set. This results in 100 tasks for annotators, where each task includes an article and six summary sentences to be evaluated—one of which originates from the reference summary and the other five are from any of the system summaries. Each task is completed by an average of 4 workers. All workers are required to have the "Master" qualification, a designation for high-quality annotations. Of the 600 summary sentences evaluated, each state-of-the-art system contributes as follows—*Bottom-Up*: 146, *DCA*: 130, *PG*: 37, *Novel*: 171, *Fast-Abs-RL*: 16, and *Reference*: 100. The number of sentences we evaluate for each system is proportional to the number of observed fusion cases.

In order to answer the *Method of Merging* and *Coverage* questions, the annotator must be provided with which two article sentences were fused together to create the summary sentence in question. We use the heuristic proposed by Lebanoff et al. (2019) to estimate which pair of sentences should be chosen. They use averaged ROUGE-1, -2, -L scores (Lin, 2004) to represent sentence similarity. The heuristic calculates the ROUGE similarity between the summary sentence and each article sentence. The article sentence with the highest similarity is chosen as the first sentence, then overlapping words are removed from the summary sentence. It continues to find the article sentence most similar to the remaining summary sentence, which is chosen as the second sentence. Our interface automatically highlights this pair of sentences (Figure 1).

The same heuristic is also employed in deciding whether a summary sentence was generated by sentence compression or fusion. The algorithm halts if no article sentence is found that shares two or more content words with the summary sentence. If it halts after only one sentence is found, then it is classified as *compression*. If it finds a second sentence, then it is classified as *fusion*.

3 Results

We present experimental results in Table 2. Our findings suggest that system summary sentences formed by fusion have low faithfulness (61.7% on average) as compared to the reference summaries. This demonstrates the need for current summarization models to put more emphasis on improving the faithfulness of generated summaries. Surprisingly, the highest performing systems, DCA and Bottom-Up, according to ROUGE result in

System	Faithful	Grammatical	Coverage
DCA	47.0	72.4	62.6
Bottom-Up	56.9	78.9	78.5
Novel	58.5	78.5	75.3
Fast-Abs-RL	69.0	77.6	82.8
PG	76.9	84.6	89.5
Reference	88.4	91.6	74.9

Table 2: Percentage of summary sentences that are faithful, grammatical, etc. according to human evaluation of several state-of-the-art summarization systems (see §2 for details).

the lowest scores for being faithful to the article. While we cannot attribute the drop in faithfulness to an over-emphasis on optimizing automatic metrics, we can state that higher ROUGE scores does not necessarily lead to more faithful summaries, as other works have shown (Falke et al., 2019). Bottom-Up, interestingly, is 20 points lower than PG, which it is closely based on. It uses an external content selector to choose what words to copy from the article. While identifying summary-worthy content improved ROUGE, we believe that Bottom-Up stitches together sections of content that do not necessarily belong together. Thus, it is important to identify not just summary-worthy content, but also *mergeable* content.

System summary sentences created by fusion are generally grammatical (78.4% on average), though it is still not up to par with reference summaries (91.6%). The chosen state-of-the-art systems use the encoder-decoder architecture, which employs a neural language model as the decoder, and language models generally succeed at encoding grammar rules and staying fluent (Clark et al., 2019). The coverage for reference summaries is moderately high (74.9%), demonstrating the effectiveness of the heuristic of identifying where summary content is pulled from. Especially for most of the systems, the heuristic successfully finds the correct source sentences. As it is based mostly on word overlap, the heuristic works better on summaries that are more extractive, hence the higher coverage scores among the systems compared to reference summaries, which are more abstractive.

Figure 2 illustrates the frequency of each merging method over the summarization systems. Most summary sentences are formed by concatenation. PG in particular most often fuses two sentences using concatenation. Surprisingly, very few reference summaries use entity replacement when performing fusion. We believe this is due to the extractiveness of the CNN/DailyMail dataset, and would likely have higher occurrences in more abstractive datasets.

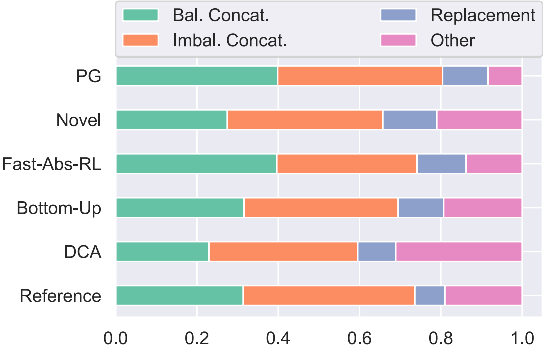


Figure 2: Frequency of each merging method. Concatenation is the most common method of merging.

System	Faithful	Grammatical	Coverage
Bal Concat	82.55	86.91	94.43
Imbal Concat	69.40	80.25	84.58
Replacement	53.06	82.04	77.55
Other	25.20	68.23	27.04

Table 3: Results for each merging method. Concatenation has high faithfulness, grammaticality, and coverage, while Replacement and Other have much lower scores.

Does the way sentences are fused affect their faithfulness and grammaticality? Table 3 provides insights regarding this question. Grammaticality is relatively high for all merging categories. Coverage is also high for balanced/imbalanced concatenation and replacement, meaning the heuristic works succesfully for these forms of sentence merging. It does not perform as well on the Other category. This is understandable, since sentences formed in a more complex manner will be harder to identify using simple word overlap. Faithfulness has a similar trend, with summaries generated using concatenation being more likely to be faithful to the original article. This may explain why PG is the most faithful of the systems, while being the simplest—it uses concatenation more than any of the other systems. We believe more effort can be directed towards improving the more complex merging paradigms, such as entity replacement.

There are a few potential limitations associated with the experimental design. Judging whether a sentence is faithful to the original article can be a difficult task to perform reliably, even for humans. We observe that the reference summaries achieve lower than the expected faithfulness and grammaticality of 100%. This can have two reasons. First, the inter-annotator agreement for this task is relatively low and we counteract this by employing an average of four annotators to complete each task. Second, we make use of an automatic heuristic to highlight sentence pairs from the article. While it generally finds the correct sentences—average Coverage score of 77.3%—the incorrect pairs may have biased the annotators away from sentences that humans would have found more appropriate. This further exemplifies the difficulty of the task.

4 Related Work

Sentence fusion aims to produce a single summary sentence by fusing multiple source sentences. Dependency graphs and discourse structure have proven useful for aligning and combining multiple sentences into a single sentence (Barzilay and McKeown, 2005; Marsi and Krahmer, 2005; Filippova and Strube, 2008; Cheung and Penn, 2014; Gerani et al., 2014). Mehdad et al. (2013) construct an entailment graph over sentences for sentence selection, then fuse sentences together using a word graph. Abstract meaning representation and other graph-based representations have also shown success in sentence fusion (Liu et al., 2015; Nayeem et al., 2018). Geva et al. (2019) fuse pairs of sentences together using Transformer, focusing on discourse connectives between sentences.

Recent summarization research has put special emphasis on faithfulness to the original text. Cao et al. (2018a) use seq-to-seq models to rewrite templates that are prone to including irrelevant entities. Incorporating additional information into a seq-to-seq model, such as entailment and dependency structure, has proven successful (Li et al., 2018; Song et al., 2018). The closest work to our human evaluation seems to be from Falke et al. (2019). Similar to our work, they find that the PG model is more faithful than Fast-Abs-RL and Bottom-Up, even though it has lower ROUGE. They show that 25% of outputs from these state-of-the-art summarization models are unfaithful to the original article. Cao et al. (2018b) reveal a similar finding that 27% of the summaries generated by a neural sequence-to-sequence model have errors. Our study, by contrast, finds 38% to be unfaithful, but we limit our study to only summary sentences created by *fusion*. Our work examines a wide variety of state-of-the-art summarization systems, and perform in-depth analysis over other measures including grammaticality, coverage, and method of merging.

5 Conclusion

In this paper we present an investigation into sentence fusion for abstractive summarization. Several state-of-the-art systems are evaluated, and we find that many of the summary outputs generate false information. Most of the false outputs were generated by entity replacement and other complex merging methods. These results demonstrate the need for more attention to be focused on improving sentence fusion and entity replacement.

Acknowledgments

We are grateful to the anonymous reviewers for their helpful comments and suggestions. This research was supported in part by the National Science Foundation grant IIS-1909603.

References

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, 31(3):297–328.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018a. Retrieve, Rerank and Rewrite: Soft Template Based Neural Summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 152–161, Melbourne, Australia. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018b. Faithful to the original: Fact aware neural abstractive summarization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep Communicating Agents for Abstractive Summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1662–1675, New Orleans, Louisiana. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Jackie Chi Kit Cheung and Gerald Penn. 2014. Unsupervised Sentence Enhancement for Automatic Summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 775–786, Doha, Qatar. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 137–144, Manchester, UK. Coling 2008 Organizing Committee.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence Compression by Deletion with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368, Lisbon, Portugal. Association for Computational Linguistics.

Katja Filippova and Michael Strube. 2008. Sentence fusion via dependency graph compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*, page 177, Honolulu, Hawaii. Association for Computational Linguistics.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-Up Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613, Doha, Qatar. Association for Computational Linguistics.

Mor Geva, Eric Malmi, Idan Szpektor, and Jonathan Berant. 2019. DiscoFuse: A Large-Scale Dataset for Discourse-Based Sentence Fusion. *arXiv:1902.10526 [cs]*. ArXiv: 1902.10526.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Proceedings of Neural Information Processing Systems (NIPS)*.

Wojciech Kryciski, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving Abstraction in Text Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. 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.

Chen Li, Fei Liu, Fuliang Weng, and Yang Liu. 2013. Document summarization via guided sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 490–500, Seattle, Washington, USA. Association for Computational Linguistics.

Chen Li, Yang Liu, Fei Liu, Lin Zhao, and Fuliang Weng. 2014. Improving multi-documents summarization by sentence compression based on expanded constituent parse trees. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 691–701, Doha, Qatar. Association for Computational Linguistics.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the Correctness of the Summary: Incorporate Entailment Knowledge into Abstractive Sentence Summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1430–1441, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A. Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Denver, Colorado. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. 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)*.

Ryan McDonald. 2006. Discriminative sentence compression with soft syntactic evidence. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. 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.

Yashar Mehdad, Giuseppe Carenini, Frank Wm Tompa, and Raymond T. Ng. 2013. Abstractive Meeting Summarization with Entailment and Fusion. In *ENLG*.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. Abstractive Unsupervised Multi-Document Summarization using Paraphrastic Sentence Fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Kaiqiang Song, Lin Zhao, and Fei Liu. 2018. Structure-Infused Copy Mechanisms for Abstractive Summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1717–1729, Santa Fe, New Mexico, 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.

Lu Wang, Hema Raghavan, Vittorio Castelli, Radu Florian, and Claire Cardie. 2013. A sentence compression based framework to query-focused multi-document summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1384–1394, Sofia, Bulgaria. Association for Computational Linguistics.