Learning to Fuse Sentences with Transformers for Summarization

Logan Lebanoff[†] **Franck Dernoncourt**[§]
Doo Soon Kim[§] **Lidan Wang**[§] **Walter Chang**[§] **Fei Liu**[†]

[†]University of Central Florida [§]Adobe Research

loganlebanoff@knights.ucf.edu feiliu@cs.ucf.edu
{dernonco,dkim,lidwang,wachang}@adobe.com

Abstract

The ability to fuse sentences is highly attractive for summarization systems because it is an essential step to produce succinct abstracts. However, to date, summarizers can fail on fusing sentences. They tend to produce few summary sentences by fusion or generate incorrect fusions that lead the summary to fail to retain the original meaning. In this paper, we explore the ability of Transformers to fuse sentences and propose novel algorithms to enhance their ability to perform sentence fusion by leveraging the knowledge of *points of correspondence* between sentences. Through extensive experiments, we investigate the effects of different design choices on Transformer's performance. Our findings highlight the importance of modeling points of correspondence between sentences for effective sentence fusion.

1 Introduction

A renewed emphasis must be placed on sentence fusion in the context of neural abstractive summarization. A majority of the systems are trained end-to-end (See et al., 2017; Paulus et al., 2018; Narayan et al., 2018; Chen and Bansal, 2018; Gehrmann et al., 2018; Liu and Lapata, 2019), where an abstractive summarizer is rewarded for generating summaries that contain the same words as human abstracts, measured by automatic metrics such as ROUGE (Lin, 2004). A summarizer, however, is not rewarded for correctly fusing sentences. In fact, when examined more closely, only few sentences in system abstracts are generated by fusion (Falke et al., 2019; Lebanoff et al., 2019). For instance, 6% of summary sentences generated by Pointer-Gen (See et al., 2017) are through fusion, whereas human abstracts contain 32% fusion sentences. Moreover, sentences generated by fusion are prone to errors. They can be ungrammatical, nonsensical, or otherwise ill-formed. There is thus an urgent need to develop neural abstractive summarizers to fuse sentences properly.

The importance of sentence fusion has long been recognized by the community before the era of neural text summarization. The pioneering work of Barzilay et al. (1999) introduces an information fusion algorithm that combines similar elements across related text to generate a succinct summary. Later work, such as (Marsi and Krahmer, 2005; Filippova and Strube, 2008; Elsner and Santhanam, 2011; Thadani and McKeown, 2013; Mehdad et al., 2013), builds a dependency or word graph by combining syntactic trees of similar sentences, then employs integer linear programming to decode a summary sentence from the graph. Most of these studies have assumed a set of *similar* sentences as input, where fusion is necessary to reduce repetition. Nonetheless, humans do not limit themselves to combine similar sentences. In this paper, we pay particular attention to fuse *disparate* sentences that contain fundamentally different content but remain related to make fusion sensible (Elsner and Santhanam, 2011). In Figure 1, we provide an example of a sentence fusion instance.

We address the challenge of fusing disparate sentences by enhancing the Transformer architecture (Vaswani et al., 2017) with *points of correspondence* between sentences, which are devices that tie two sentences together into a coherent text. The task of sentence fusion involves choosing content from each sentence and weaving the content pieces together into an output sentence that is linguistically plausible and semantically truthful to the original input. It is distinct from Geva et al. (2019) that connect two sentences with discourse markers. Our contributions are as follows.

- We make crucial use of *points of correspondence* (PoC) between sentences for information fusion. Our use of PoC was initiated by the current lack

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4136–4142,
November 16–20, 2020.

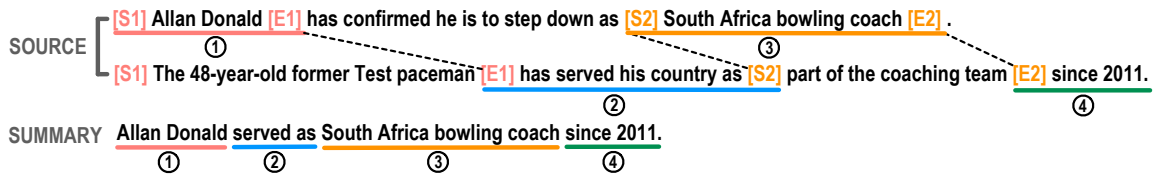

Figure 1: Sentence fusion involves determining what content from each sentence to retain, and how best to weave text pieces together into a well-formed sentence. Points of correspondence (PoC) are text chunks that convey the same or similar meanings, e.g., *Allan Donald* and *The 48-year-old former Test paceman*, *South Africa bowling coach* and *part of the coaching team*.

of understanding of how sentences are combined in neural text summarization.

- We design new sentence fusion systems and experiment with a fusion dataset containing quality PoC annotations as the test bed for this investigation. Our findings highlight the importance of modeling points of correspondence for fusion.[1]

2 Method

A PoC is a pair of text chunks that express the same or similar meanings. In Fig. 1, *Allan Donald* vs. *The 48-year-old former Test paceman*, *South Africa bowling coach* vs. *part of the coaching team* are two PoCs. The use of alternative expressions for conveying the same meanings is standard practice in writing, as it increases lexical variety and reduces redundancy. However, existing summarizers cannot make effective use of these expressions to establish correspondence between sentences, often leading to ungrammatical and nonsensical outputs.

2.1 Transformer with Linking

It is advantageous for a Transformer model to make use of PoC information for sentence fusion. While Transformer-based pretrained models have had considerable success (Devlin et al., 2019; Dong et al., 2019; Lewis et al., 2020), they primarily feature pairwise relationships between *tokens*, but not PoC mentions, which are are *text chunks* of varying size. Only to a limited extent do these models embed knowledge of coreference (Clark et al., 2019), and there is a growing need for incorporating PoC linkages explicitly in a Transformer model to enhance its ability to perform sentence fusion.

We propose to enrich Transformer's source sequence with *markups* that indicate PoC linkages. Here PoC information is assumed to be available for any fusion instance (details in §3). We introduce special tokens ([S$_k$] and [E$_k$]) to mark the start and end of each PoC mention; all mentions pertaining to the k-th PoC share the same start/end tokens. An example is illustrated in Figure 1, where *Allan Donald* and *The 48-year-old former Test paceman* are enriched with the same special tokens. We expect special tokens to assist in linking coreferring mentions, creating long-range dependencies between them and encouraging the model to use these mentions interchangeably in generation (Figure 2). The model is called "TRANS-LINKING."

Our Transformer takes as input a sequence $\mathcal{S}$ formed by concatenating the source and summary sequences. Let $\mathbf{H}^l = [\mathbf{h}_1^l, \ldots, \mathbf{h}_{|\mathcal{S}|}^l]$ be hidden representations of the l-th layer of a decoder-only architecture. An attention head transforms each vector respectively into query ($\mathbf{q}_i$), key ($\mathbf{k}_j$) and value ($\mathbf{v}_j$) vectors. The attention weight $\alpha_{i,j}$ is computed for all pairs of tokens by taking the scaled dot product of query and key vectors and applying softmax over the output (Eq. (1)). $\alpha_{i,j}$ indicates the importance of token j to constructing $\mathbf{h}_i^l$ of the current token i.

$$\alpha_{i,j} = \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_j / \sqrt{d_k} + \mathcal{M}_{i,j})}{\sum_{j'=1}^{|\mathcal{S}|} \exp(\mathbf{q}_i^\top \mathbf{k}_{j'} / \sqrt{d_k} + \mathcal{M}_{i,j'})} \quad (1)$$

We utilize a mask $\mathcal{M} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ to control the attention of the model (Eq. (2)). $\mathcal{M}_{i,j} = 0$ allows token i to attend to j and $\mathcal{M}_{i,j} = -\infty$ prevents i from attending to j as it leads $\alpha_{i,j}$ to be zero after softmax normalization. Similar to (Dong et al., 2019), a source token ($i \leq |\mathbf{x}|$) can attend to all other source tokens ($\mathcal{M}_{i,j} = 0$ for $j \leq |\mathbf{x}|$). A summary token ($i > |\mathbf{x}|$) can attend to all tokens including itself and those prior to it ($\mathcal{M}_{i,j} = 0$ for $j \leq i$). The mask $\mathcal{M}$ provides desired flexibility in terms of building hidden representations for tokens in $\mathcal{S}$. The output of the attention head is a weighted sum of the value vectors $\mathbf{h}_i^l = \sum_{j=1}^{|\mathcal{S}|} \alpha_{i,j} \mathbf{v}_j$.

$$\mathcal{M}_{i,j} = \begin{cases} 0 & \text{if } j \leq \max(i, |\mathbf{x}|) \\ -\infty & \text{otherwise} \end{cases} \quad (2)$$

[1] Our code is publicly available at https://github.com/ucfnlp/sent-fusion-transformers

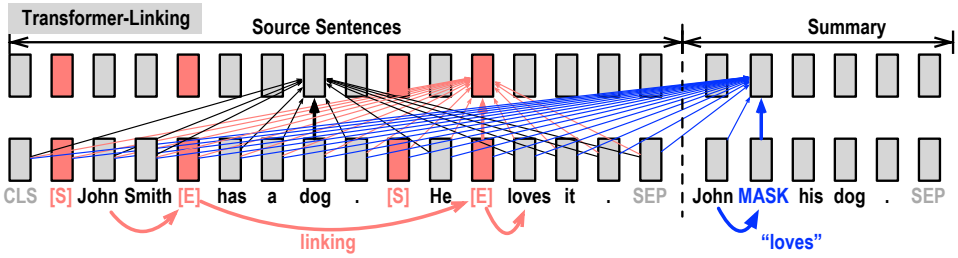

Figure 2: Our TRANS-LINKING model facilitates summary generation by reducing the shifting distance, allowing the model attention to shift from "John" to the tokens "[E]" then to "loves" for predicting the next summary word.

We fine-tune the model on a sentence fusion dataset (§3) using a denoising objective, where 70% of the summary tokens are randomly masked out. The model is trained to predict the original tokens conditioned on hidden vectors of MASK tokens: $\mathbf{o} = \text{softmax}(\mathbf{W}^O \text{GeLU}(\mathbf{W}^h \mathbf{h}^L_{\text{MASK}}))$, where parameters $\mathbf{W}^O$ are tied with token embeddings. By inserting markup tokens, our model provides a soft linking mechanism to allow mentions of the same PoC to be used interchangeably in summary generation. As shown in Figure 2, without PoC linking, the focus of the model attention has to shift a long distance from "John" to "loves" to generate the next summary word. Their long-range dependency is not always effectively captured by the model. In contrast, our TRANS-LINKING model substantially reduces the shifting distance, allowing the model to hop to the special token "[E]" then to "loves," facilitating summary generation.

2.2 Transformer with Shared Representation

We explore an alternative method to allow mentions of the same PoC to be connected with each other. Particularly, we direct one attention head to focus on tokens belonging to the same PoC, allowing these tokens to share semantic representations, similar to Strubell et al. (2018). Sharing representation is meaningful as these mentions are related by complex morpho-syntactic, syntactic or semantic constraints (Grosz et al., 1995).

Let $\mathbf{z}=\{z_1, \ldots, z_{|\mathbf{z}|}\}$ be a sequence containing PoC information, where $z_i \in \{0, \ldots, \mathsf{K}\}$ indicates the index of PoC to which the token $\mathbf{x}_i$ belongs. z_i=0 indicates $\mathbf{x}_i$ is not associated with any PoC. Our TRANS-SHAREREPR model selects an attention head h from the l-th layer of the Transformer model. The attention head h governs tokens that belong to PoCs ($z_i \neq 0$). Its hidden representation $\mathbf{h}^l_i$ is computed by modeling only pairwise relationships between token i and any token j of the same PoC ($z_i = z_j$; Eq. (3)), while other tokens

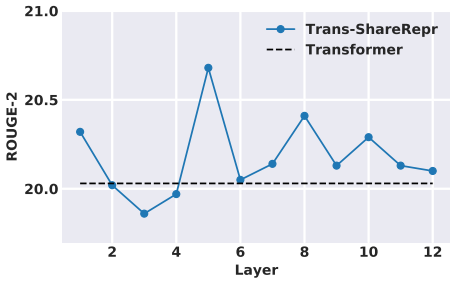

Figure 3: The first attention head from the l-th layer is dedicated to coreferring mentions. The head encourages tokens of the same PoC to share similar representations. Our results suggest that the attention head of the 5-th layer achieves competitive performance, while most heads perform better than the baseline. The findings are congruent with (Clark et al., 2019) that provides a detailed analysis of BERT's attention.

are excluded from consideration.

$$\mathcal{M}^h_{i,j} = \begin{cases} 0 & \text{if } i, j \leq |\mathbf{x}| \;\&\; z_i = z_j \\ -\infty & \text{otherwise} \end{cases} \quad (3)$$

For example, "*Allan Donald*" and "*The 48-year-old former Test paceman*" are co-referring mentions. TRANS-SHAREREPR allows these tokens to only attend to each other when learning representations using the attention head h. These tokens are likely to yield similar representations. The method thus accomplishes a similar goal as TRANS-LINKING to allow *tokens* of the same PoC to be treated equivalently during summary generation; we explore the selection of attention heads in §3.

3 Experiments

Corpus Our corpus contains a collection of documents, source and fusion sentences, and human annotations of corresponding regions between sentences. The set of documents were sampled from CNN/DM (See et al., 2017) and PoC annotations were obtained from Lebanoff et al. (2020). They use a human summary sentence as an anchor point

System	Heuristic Set				Point of Correspondence Test Set						
	R-1	R-2	R-L	BLEU	R-1	R-2	R-L	BLEU	B-Score	#Tkns	%Fuse
Pointer-Generator	35.8	18.2	31.8	41.9	33.7	16.3	29.3	40.3	57.3	14.3	38.7
Transformer	39.6	20.9	35.3	47.2	38.8	20.0	33.8	45.8	61.3	15.1	50.7
Trans-LINKING	**39.8**	**21.1**	**35.3**	**47.3**	38.8	20.1	33.9	45.5	61.1	15.1	**55.8**
Trans-SHAREREPR	39.4	20.9	35.2	46.9	**39.0**	**20.2**	**33.9**	**45.8**	61.2	**15.2**	46.5
Concat-Baseline	37.2	20.0	28.7	25.0	36.1	18.6	27.8	24.6	60.4	52.0	99.7

Table 1: Results of various sentence fusion systems. We report the percentage of output sentences that are generated by fusion (%Fuse) and the average number of tokens per output sentence (#Tkns). To calculate %Fuse, we follow the same procedure used by Lebanoff et al. (2020) – a generated sentence is regarded as a fusion if it contains at least two non-stopword tokens from each sentence that do not already exist in the other sentence.

to find two document sentences that are most similar to it, which forms a fusion instance containing a pair of source sentences and their summary. PoCs have been annotated based on Halliday and Hasan's theory of cohesion (1976) for 1,494 fusion instances, taken from 1,174 documents in the test and valid splits of CNN/DM with a moderate to high inter-annotator agreement (0.58).

Automatic Evaluation We proceed by investigating the effectiveness of various sentence fusion models, including (a) **Pointer-Generator** (See et al., 2017) that employs an encoder-decoder architecture to condense input sentences to a vector representation, then decode it into a fusion sentence. (b) **Transformer**, our baseline Transformer architecture w/o PoC information. It is a strong baseline that resembles the UniLM model described in (Dong et al., 2019). (c) **TRANS-LINKING** uses special tokens to mark the boundaries of PoC mentions (§2.1). (d) **TRANS-SHAREREPR** allows tokens of the same PoC to share representations (§2.2). All Transformer models are initialized with BERT-BASE parameters and are fine-tuned using UniLM's sequence-to-sequence objective for 11 epochs, with a batch size of 32. The source and fusion sentences use BPE tokenization, and the combined input/output sequence is truncated to 128 tokens. We use the Adam optimizer with a learning rate of 2e-5 with warm-up. For PG, we use the default settings and truncate the output sequences to 60 tokens.

All of the fusion models are trained (or fine-tuned) on the same training set containing 107k fusion instances from the training split of CNN/DM; PoC are identified by the spaCy coreference resolver. We evaluate fusion models on two test sets, including a "heuristic set" containing testing instances and automatically identified PoC via spaCy, and a final test set containing 1,494 instances with human-labelled PoC. We evaluate only on the instances that contain at least one point of correspondence, so we have to disregard a small percentage of instances (6.6%) in the dataset of Lebanoff et al. (2020) that contain no points of correspondence.

Source: Later that month, the ICC opened a preliminary examination into the situation in Palestinian territories, paving the way for possible war crimes investigations against Israelis.
Israel and the United States, neither of which is an ICC member, opposed the Palestinians' efforts to join the body.

Pointer-Generator: *ICC opened a preliminary examination into the situation in Palestinian territories .*

Transformer: *Israel, U.S. and the United States are investigating possible war crimes, paving way for war crimes.*

Transformer-SHAREREPR: *Israel and U.S. opposed the ICC's investigation into the situation in Palestinian territories.*

Reference: *Israel and the United States opposed the move, which could open the door to war crimes investigations against Israelis.*

Table 2: Example output of sentence fusion systems. **PG** only performs sentence shortening rather than fusion. **Transformer** fails to retain the original meaning and **Transformer-SHAREREPR** performs best. **Reference** demonstrates a high level of abstraction. Sentences are manually de-tokenized for readability.

We compare system outputs and references using a number of automatic evaluation metrics including ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020). Results are presented in Table 1. We observe that all Transformer models outperform PG, suggesting that these models can benefit substantially from unsupervised pretraining on a large corpus of text. On the heuristic test set where training and testing conditions match (they both use automatically identified PoC), **TRANS-LINKING** performs better than **TRANS-SHAREREPR**, and vice versa on the final test set. We conjecture that this is because the linking model has a stronger requirement on PoC boundaries and the training/testing conditions must match for it to be effective. In contrast, **TRANS-SHAREREPR** is more lenient with mismatched conditions.

We include a **CONCAT-BASELINE** that creates a fusion by simply concatenating two input sentences. Its

output contains 52 tokens on average, while other model outputs contain 15 tokens. This is a 70% compression rate, which adds to the challenge of content selection (Daume III and Marcu, 2004). Despite that all models are trained to fuse sentences, their outputs are not guaranteed to be fusions and shortening of single sentences is possible. We observe that **Trans-Linking** has the highest rate of producing fusions (56%). In Figure 3, we examine the effect of different design choices, where the first attention head of the l-th layer is dedicated to PoC. We report the averaged results in Table 1.

Human evaluation We investigate the quality of fusions with human evaluation. The models we use for comparison include (a) **Pointer-Generator**, (b) **Transformer**, (c) **Trans-ShareRepr** and (d) human reference fusion sentences. Example outputs for each model can be seen in Table 2. We perform evaluation on 200 randomly sampled instances from the point of correspondence test set. We take an extra step to ensure all model outputs for selected instances contain fusion sentences, as opposed to shortening of single sentences. A human evaluator from Amazon Mechanical Turk (`mturk.com`) is asked to assess if the fusion sentence has successfully retained the original meaning. Specifically, an evaluator is tasked with reading the two article sentences and fusion sentence and answering yes or no to the following question, "Is this summary sentence true to the original article sentences it's been sourced from, and it has not added any new meaning?" Each instance is judged by five human evaluators and results are shown in Table 3. Additionally, we measure their *extractiveness* by reporting on the percentage of n-grams (n=1/2/3) that appear in the source. Human sentence fusions are highly abstractive, and as the gold standard, we wish to emulate this level of abstraction in automatic summarizers. Fusing two sentences together coherently requires connective phrases and sometimes requires rephrasing parts of sentences. However, higher abstraction does not mean higher quality fusions, especially in neural models.

Interestingly, we observe that humans do not always rate reference fusions as truthful. This is in part because reference fusions exhibit a high level of abstraction and they occasionally contain content not in the source. If fusion sentences are less extractive, humans sometimes perceive that as less truthful, especially when compared to fusions that reuse the source text. Our results call for a reexamination of sentence fusion using better evaluation metrics including semantics and question-answering-based metrics (Zhao et al., 2019; Wang et al., 2020; Durmus et al., 2020).

System	Truthful.	Extractiveness 1-gram	2-gram	3-gram
Pointer-Generator	63.6	97.5	83.1	72.8
Transformer	71.7	91.9	68.6	54.2
Trans-ShareRepr	70.9	92.0	70.1	56.4
Reference	67.2	72.0	34.9	20.9

Table 3: Fusion sentences are evaluated by their level of truthfulness and extractivenss. Our system fusions attain a high level of truthfulness with moderate extractivenss.

4 Conclusion

We address the challenge of information fusion in the context of neural abstractive summarization by making crucial use of points of correspondence between sentences. We enrich Transformers with PoC information and report model performance on a new test bed for information fusion. Our findings suggest that modeling points of correspondence is crucial for effective sentence fusion, and sentence fusion remains a challenging direction of research. Future work may explore the use of points of correspondence and sentence fusion in the standard setting of document summarization. Performing sentence fusion accurately and succinctly is especially important for summarizing long documents and book chapters (Ladhak et al., 2020). These domains may contain more entities and events to potentially confuse a summarizer, making our method of explicitly marking these entities beneficial.

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

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.

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.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13063–13075. Curran Associates, Inc.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

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.

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 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.

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.

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.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Michael A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. English Language Series. Longman Group Ltd.

Faisal Ladhak, Bryan Li, Yaser Al-Onaizan, and Kathleen McKeown. 2020. Exploring content selection in summarization of novel chapters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5043–5054, Online. Association for Computational Linguistics.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019. Analyzing sentence fusion in abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 104–110, Hong Kong, China. Association for Computational Linguistics.

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*, Seattle, United States. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

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.

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)*.

Yashar Mehdad, Giuseppe Carenini, Frank Tompa, and Raymond T. NG. 2013. Abstractive meeting summarization with entailment and fusion. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 136–146, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

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.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.