ADVANCES IN KNOWLEDGE ORGANIZATION
Proceedings of International ISKO Conference - ISSN: 0938-5495 - 2026 - São Paulo

Text Summarization Generation Incorporating Factual Sequences from Social Media - A Case Study of Reddit

Ruiyang Chen

Center for Studies of Information Resources, Wuhan University, China

Lu An[1]

Center for Studies of Information Resources, Wuhan University, China

Quan Liu

School of Information Management, Wuhan University, China

Yan Shen

Institute of Data Intelligence, Wuhan University, China

Gang Li

Center for Studies of Information Resources, Wuhan University, China

Chuanming Yu

School of Information Engineering, Zhongnan University of Economics and Law, China

Abstract

To address information overload, social media platforms try to filter or refine user-generated content. For example, Reddit provides a TL;DR (Too Long; Didn't Read) module to help users quickly understand long posts or comments. But TL;DR summaries often omit important context and details and may fail to preserve the factual structure of the original text. This study situates this problem within knowledge organization (KO), which is concerned with the representation, structuring, and mediation of access to knowledge resources. From this perspective, social media summarization can be understood as a process of transforming loosely structured discourse into concise and accessible knowledge representation. This study proposes a novel text summarization technique for social media platforms. Topic sentences are first identified in the source text on Reddit based on TL;DR. Then, fact sequences are extracted from the topic sentences based on dependent syntactic analysis. Finally, a text summarization model is proposed to fuse the fact sequences. Experiments on public datasets show that our method outperforms the baseline models, especially on the metrics of ROUGE-1, ROUGE-L, accuracy, fluency and informativeness. Overall, this study generates high-quality summaries and supports knowledge extraction from social media.

Keywords: Text summarization; Knowledge organization; Social media; Fact Extraction; Semantic Representation

[1] Corresponding author.

1.0 Introduction

With the popularity of social media, information overload has become a pervasive problem. Text summarization technology is one of the important technical means to alleviate information overload (Tripathi et al. 2026). This study approaches social media summarization from the perspective of knowledge organization (KO). KO is concerned with how knowledge is described, represented, and made accessible for users. In the context of social media, massive and loosely structured discourse needs to be filtered and reorganized into usable forms of information. In this sense, automatic summarization is not only a natural language generation task, but also a form of organizing unstructured information into more accessible knowledge representation. It is also consistent with the view that knowledge organization is a social and dynamic process rather than a purely static classification activity (Hjørland 2013).

Research on text summarization for social media platforms has been conducted in recent years (Martínez-Avila et al. 2023), mainly consisting of dataset construction and algorithm research. However, few studies have fully incorporated platform-specific features of reddit to improve model performance. Reddit has the function of TL;DR (Too Long; Didn't Read), which provides short summaries of posts or comments. Although TL;DR summaries ca help users quickly grasp the main idea, they may lack important background information, factual details, or semantic relations. Existing neutral summarization models can generate fluent and semantically coherent summaries, but they may overlook key information or fail to preserve the factual structure of the source text as it is included in TL;DR. From a KO perspective, this creates a problem of incomplete or insufficient knowledge representation: the summary may be readable, but it may not fully support users' understanding of the original content.

This study tries to establish the connection between TL;DR and source text. Inspired by TLDR9+ (Sotudeh et al. 2021), we calculate the topic relevance score of each sentence in the source text and select the top-ranked sentences as topic sentences. Inspired by Huang et al. (2018), we propose a fact extraction workflow based on dependency analysis to capture important components of verb phrases and extract fact sequences from topic sentences. Finally, a fusion model is constructed for social media information summary generation. From the perspective of KO, this process transforms scattered social media discourse into organized knowledge, which involves identifying core themes and structuring them into coherent factual units.

The contributions of this study are as follows. First, we design a topic sentence annotation processing method for social media information with TL;DR. Second, fact sequences in topic sentences are extracted by improving dependent syntactic analysis techniques. Third, we evaluate the proposed model using ROUGE metrics and manual evaluation. Experimental results show that the model incorporating fact sequence information outperforms baseline models.

2.0 Literature Review

The influence of information overload on users is divided into three main aspects: cognitive, emotional, and behavioral. According to the information processing theory (Atkinson and Shiffrin 1968), the cognitive processing capacity of individuals is limited. When the amount of information exceeds the processing capacity of individuals, it hinders individuals from correctly understanding information and distinguishing important information. In response to the information overload problem, coping strategies can be adopted at the individual level, the information level, and the system and platform level, respectively. In KO, this problem is also related to the longstanding goal of organizing information so that users can identify, access, and use relevant knowledge more efficiently.

Text summarization is an important tool to alleviate information overload. Earlier neural approaches included sequence-to-sequence models with attention (Mendes et al. 2019), pointer-generator networks that copy words from the source text (See et al. 2017), and extensions that incorporate topic, structural, or semantic information (Liu et al. 2019; Hardy and Vlachos 2018; Fonseca et al. 2022). For the pre-training model, BERTSum (Liu and Lapata 2019) uses the BERT encoder for the summary and random decoder initialization. Lewis et al. (2020) proposed BART, which pre-trained both encoder and decoder. Dou et al. (2021) explored the use of “guiding signals” (such as keywords, significant sentences) to further boost performance. Zhang et al. (2023) propose an aspect-augmented unsupervised opinion summarization model, which can generate a summary that expresses the attitude of most users based on textual reviews and a heterogeneous graph. Liu et al. (2022) propose a key phrase aware transformer (KPAT) for abstractive summarization, which can highlight and encode key phrases in input documents. Zhang et al. (2021) propose a fact-aware reinforced abstractive sentence summarization framework (FAR-ASS), which can improve the factual correctness of neural abstractive summarization models.

Researchers and practitioners have made progress in the field of text summarization. However, TL;DR-style summaries often lack important context and details, leading to a biased understanding of the original text. This limitation is important for KO because a summary does not merely shorten a text; it also selects, orders, and represents knowledge for subsequent use. A summarization model that preserves factual relations and contextual cues can therefore contribute to the quality and usability of organized knowledge.

3.0 Methodology

This section describes our methods in three parts: extracting topic sentences, extracting fact sequences from the topic sentences, and constructing a summarization model with fused fact sequences.

3.1 Extracting topic sentences

This section aims to model links between TL;DR and source text. We calculate the correlation between each sentence in the source text and TL;DR. First, a target tagging task is defined. A message from Reddit is divided into sentences. Then the topic importance score (y_{s_i}) of each sentence Si is calculated according to TL;DR. See Equation (1).

$$y_{s_i} = \frac{rouge2(s_i) + rougeL(s_i)}{\sum_{s_i \in D} rouge2 + rougeL} \quad (1)$$

where rouge2 denotes the number of overlaps in binary phrases between s_i and TL;DR, and rougeL denotes the number of overlaps in the longest common subsequences between s_i and TL;DR. After obtaining the topic importance score of each sentence, sentences are sorted by y_{s_i}. A threshold N can be set to select N sentences with the highest scores as topic sentences.

3.2 Fact sequence extraction

Dependency parsing aims to determine syntactic relations between words, such as subject, predicate and object. This study uses three types of relations to construct fact units: triggers, arguments, and modifiers. The trigger is usually the root of the sentence. Arguments contain "agent" and "comp". “agent” refers to the person or thing that performs the action in a sentence. The “comp” token is used to indicate a word as a complement to a verb or adjective. There are multiple subcategories of "comp" and we have selected the following four dependencies. Acomp is an adjectival phrase that functions as the complement of a verb. Ccomp is a dependent clause acting as an object for verbs and adjectives. Xcomp is a non-finite clausal complement without its own subject, determined by an external argument. Dobj is a noun phrase that receives the action of the verb in the accusative case. The modifier refers to a word that modifies a central word. We have selected prt and neg as two common types of dependencies. Prt (phrasal verb particle) relation identifies a phrasal verb, and holds between the verb and its particle. Neg (negation modifier) is a dependency type indicating negation. Negation is a word used to express the relation between a negation word and the word it modifies.

Juxtaposition can express the relationship between multiple actions or states, better reflecting the completeness of the fact. After determining the required syntactic relations, we designed a set of matching rules to extract the fact units before constructing the fact sequence according to its syntactic relations with lexical matching and other rules.

After performing topic importance calculations with Reddit's social media feeds and its TL;DR, we filter a certain number of topic sentences. Then facts are extracted from the topic sentences. The data structure of the facts is a set containing the triggers and arguments from the syntactic analysis. spaCy is used to parse the dependencies between words in a sentence. Then event triggers and arguments are annotated according to their dependencies. We develop a fact extractor that extracts fact units from topic sentences to act as fact plots. The approach of representing facts as verb phrases is employed. Verbs, as

anchors of sentences, can be considered fact triggers. Thus, our goal is to extract all the key actors (as fact parameters) associated with fact triggers. We construct a sequence of events denoted as S, which is a sequence of knowledge consisting of one or more factual units that are stitched together and combined.

The algorithm iterates over each sentence in the Reddit input and extracts a fact sequence E for each topic sentence O_i, it initializes an empty fact object e_i, normalized the sentence, and performs dependency parsing. From the dependency parse, the fact trigger ROOT and its position are extracted. Then we iterate over the roles (trigger, arg, mod) and check if the trigger is the head and the role is the tail in the dependency parse. If so, it extracts the (role, pr) pair and sets it in the fact object e_i. Finally, it aligns the roles by their position pr and appendse_i to the fact sequence. If there are multiple topic sentences, we extract their fact sequences, marking the beginning with [START], the end with [END], and using [SEP] to separate the sequences in the middle.

3.3 Summary generation model with fused fact sequences

3.3.1 Task Definition

We improve the pre-trained generative model BART by defining the following task in the framework of social media text summary generation for Reddit. For the TL;DR module of the Reddit platform, a generation model is constructed that exploits both the context and the topic of TL;DR. Our goal is to investigate how to generate summaries that better match the source text while tracking a given sequence of facts with a neural network model. So, we add a sequence of facts for each input social media message. See Figure 1.

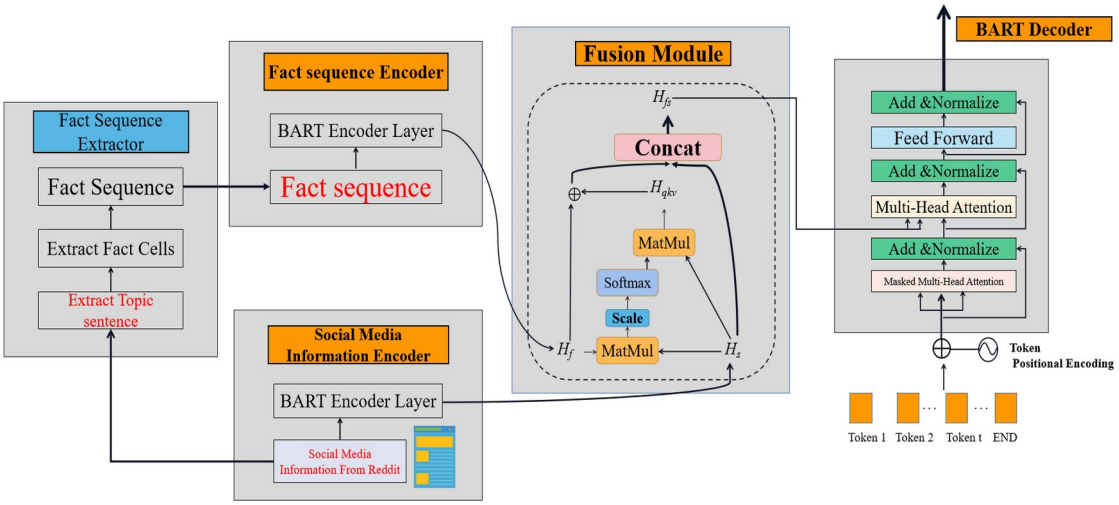

Figure 1. Diagram of the text summarization model fusing the fact sequence

3.3.2 Text summarization model

3.3.2.1 Encoding Module

We encode the social media information S through two separate channels: 1) The hidden variable Hs is obtained by encoding the social media information Encodersocial , which is constructed by the BART's Encoder; 2) The factual knowledge sequence F is obtained from the topic sentences of social media messages by the factual sequence extraction technique proposed in Section 3.2, and is similarly encoded by the factual sequence encoder Encoderfact composed of the BART's Encoder to obtain the hidden variable Hf , where both Encodersocial and Encoderfact are original BART encoders but do not share trainable parameters when fine-tuning.

3.3.2.2 Fusion Module

Inspired by cross attention mechanism (Vaswani et al 2017), we propose an information fusion technique to enhance the weight of fact sequences in social media messages. First, we construct three trainable matrices with$i$denoting the$i$-th head in the attention score and a total of *m* heads. The$i$-th head attention D_i is the attention-based weight sum of the feature matrix. Finally, the resulting H_{fs} represents the attention of the ongoing event considering the context. W_i^Q、W_i^K、W_i^V and W^m are linearly transformed weight matrices.

To enhance the contextualization of input event features, we adopt an incremental approach where we add the H_{qkv} (Contextual Feature Augmentation) to the original event features Hf. This technique enables neural models to better capture the context gap that exists between event sequences and stories. By augmenting the original features with additional contextual information, we can provide a more comprehensive understanding of the narrative and facilitate the learning of complex dependencies and relationships between facts. This allows for more accurate predictions and a deeper understanding of the underlying content. The formulas are shown in Equations (7) and (8).

$$H_{ffs} = H_f + \alpha H_{qkv} \quad (7)$$

$$H_{fs} = Concat(H_s, H_{ffs}) \quad (8)$$

where α denotes the scale factor of H_{qkv}. Finally, H_{fs} and H_{ffs} are stitched together to obtain a fused representation combining factual sequences with social media information.

3.3.2.3 Decoding process

After obtaining the fused context vector H_{fs}, we initialize the decoder with a start token y_1. At each time step t, BART's original decoder generates the output token y_t based on the previously generated token $y_{t-1}$and the context vector. The decoder uses a self-attentive mechanism at each time step to compute the output at the current moment while using a multi-headed attention mechanism to weight the input sequence to obtain contextual information. The next token is sampled as the output from the generated probability distribution $p(\{y_t | y_{<t}\}, c)$, where c denotes the input to the neural network model. During the decoding process, it is also necessary to consider the conditions for stopping the generation. At each time step t, we need to check if the generated token is an end token or if the maximum generation length has been reached.

4.0 Experiment

4.1 Dataset

The TLDRHQ dataset is composed of long texts and corresponding summaries extracted from web pages in different domains and sources. We randomly select 412,763 entries in the TLDRHQ dataset as our experimental data. The data distribution is as follows: training set (393,207 entries), validation set (9,821 entries), and test set (9,735 entries).

4.2 Baseline methods

The baseline methods include Seq2seq+Attention (Bahdanau et al. 2015), Pointer Network (See et al. 2017), CopyNet (Gu et al. 2016), S2S+ATT (copy+coverage), GPT2 (Radford et al. 2019), PEGASUS (Zhang et al. 2020) and BART (Lewis et al. 2020).

4.3 Automatic evaluation of metrics

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is an automatic abstract evaluation method proposed by (Lin 2004). It compares an automatically generated abstract or translation with a set of reference abstracts and counts the number of overlapping basic units between them to derive a corresponding score. The three most commonly used ROUGE metrics are ROUGE-1 (the overlap of 1-gram between the summary

generated by a system and reference summaries), ROUGE-2 (bi-grams) and ROUGE-L (the longest common subsequence).

4.4 Implementation process

BART and our model use the BART-based pretrained model released by Facebook. The learning rate is 2e-4 for BART fine-tuning and 6e-5 for our model. GPT2 and PEGASUS are fine-tuned with a learning rate of 1e-4, while the non-pretrained baselines use the same seq2seq framework with different attention, copy, and coverage settings.

4.5 Performance of experiments

Table 1 shows that our model outperforms most of the baseline models, especially on Rouge-1 and Rouge-L.

Table 1 Performance of Different Models

Models	ROUGE-1	ROUGE-2	ROUGE-L
Non-pretrained Language Model			
Seq2seq+Attention	20.35	7.40	15.63
CopyNet	20.98	8.13	16.45
PointerNet	21.42	8.10	16.87
S2S+ATT (copy+coverage)	22.70	9.35	18.61
Pretraining Language Models			
BART	27.62	9.29	21.40
PEGASUS	31.40	**14.09**	24.61
GPT2	27.11	8.61	20.93
OURS	**32.40**	12.90	**25.37**

4.6 Ablation experiments

We replaced the fusion module in the model with vector stitching and explored the effect of the fusion module on experimental performance by comparing the experiments in different ways. The experimental results are shown in Table 2.

Table 2 Comparative Experimental Results of Models with and without the Fusion Module

Models	Rouge-1	Rouge-2	Rouge-L
Ours (concat)	31.02	11.24	24.45
Ours (fusion module)	**32.40**	**12.90**	**25.37**

It is seen that the fusion module is significantly more effective than vector stitching. With the fusion module, each head can focus on different features and capture the interaction features between two vectors, thus achieving fine-tuning of different information and accurate fusion. In contrast, vector stitching cannot process different dimensions of information separately and cannot learn the interaction features. Table 3 shows the effect of the number of topic sentences on the experimental results. It is seen that the effect is close when one or two topic sentences are selected. The performance starts to decrease when three topic sentences are selected.

Table 3 Effect of the Number of Topic Sentences on the Experimental Result

Models	Rouge-1	Rouge-2	Rouge-L
Ours (n=1)	31.98	11.56	25.11
Ours (n=2)	**32.40**	**12.90**	**25.37**
Ours (n=3)	29.63	10.23	23.13

4.7 Manual evaluation

Twenty test samples were randomly selected and evaluated for readability, accuracy, fluency, and informativeness on a 1-5 scale. Compared with three pre-trained models, our model performs the best in terms of accuracy, fluency and informativeness, indicating that our model can produce high-quality text and convey rich information as shown in Figure 2.

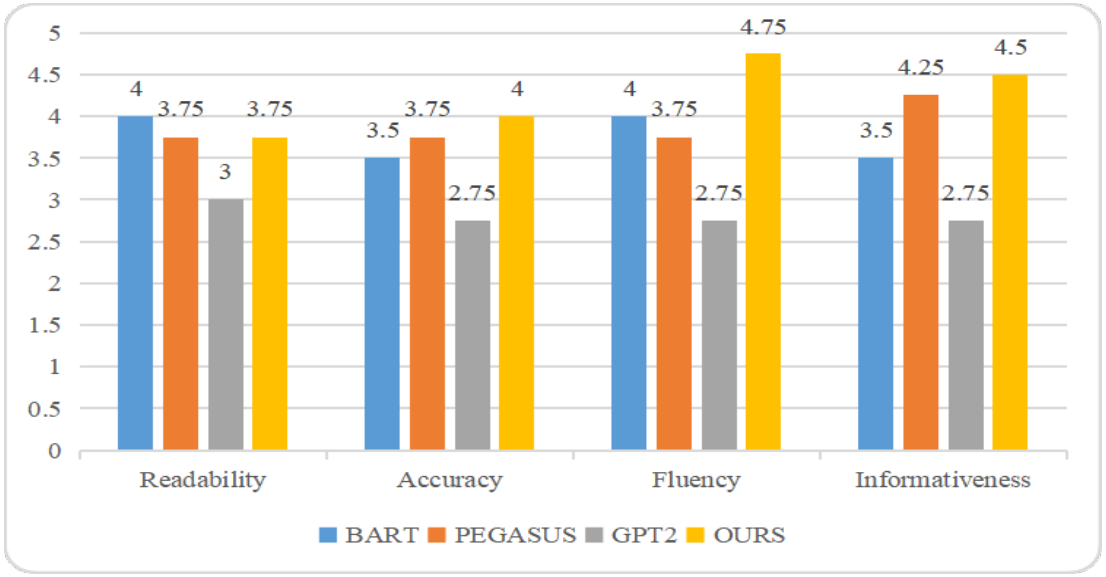

Figure 2. Manual evaluation results

5.0 Conclusion

The problem of information overload on social media platforms is becoming increasingly prevalent. To address the limitations and gaps in text summarization, we propose a novel text summarization technique for the Reddit platform. Our approach first identifies topic sentences in the source text based on TL;DR. Then fact sequences are extracted from these topic sentences using dependent syntactic analysis. Finally, a text summarization model is proposed that combines these fact sequences. Our method outperforms the baseline model on publicly available datasets, demonstrating its effectiveness.

Acknowledgements: This work was supported by the National Social Science Foundation of China (Grant No. 23&ZD230).

AI Usage Statement: The authors declare that no generative artificial intelligence (AI) tools were used in the research.

References

Atkinson, R. C., and R. M. Shiffrin. 1968. “Human Memory: A Proposed System and Its Control Processes.” In The Psychology of Learning and Motivation: Advances in Research and Theory, vol. 2, edited by Kenneth W. Spence and Janet T. Spence, 89-195. New York: Academic Press. doi:10.1016/S0079-7421(08)60422-3.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. “Neural Machine Translation by Jointly Learning to Align and Translate.” In 3rd International Conference on

Learning Representations (ICLR 2015), Conference Track Proceedings, edited by Yoshua Bengio and Yann LeCun. San Diego, CA. arXiv:1409.0473.

Dou, Zi-Yi, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. "GSum: A General Framework for Guided Neural Abstractive Summarization." In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, edited by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, 4830-42. Online: Association for Computational Linguistics. doi:10.18653/v1/2021.naacl-main.384.

Fonseca, Marcio, Yftah Ziser, and Shay B. Cohen. 2022. "Factorizing Content and Budget Decisions in Abstractive Summarization of Long Documents." In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, edited by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, 6341-64. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. doi:10.18653/v1/2022.emnlp-main.426.

Gu, Jiatao, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. "Incorporating Copying Mechanism in Sequence-to-Sequence Learning." In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), edited by Katrin Erk and Noah A. Smith, 1631-40. Berlin: Association for Computational Linguistics. doi:10.18653/v1/P16-1154.

Hardy, and Andreas Vlachos. 2018. "Guided Neural Language Generation for Abstractive Summarization Using Abstract Meaning Representation." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, edited by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, 768-73. Brussels: Association for Computational Linguistics. doi:10.18653/v1/D18-1086.

Hjørland, Birger. 2013. "Citation Analysis: A Social and Dynamic Approach to Knowledge Organization." Information Processing & Management 49: 1313-25. doi:10.1016/j.ipm.2013.07.001.

Huang, Lifu, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. "Zero-Shot Transfer Learning for Event Extraction." In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), edited by Iryna Gurevych and Yusuke Miyao, 2160-70. Melbourne: Association for Computational Linguistics. doi:10.18653/v1/P18-1201.

Lewis, Mike, 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, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 7871-80. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.703.

Lin, Chin-Yew. 2004. “ROUGE: A Package for Automatic Evaluation of Summaries.” In Text Summarization Branches Out, 74-81. Barcelona: Association for Computational Linguistics.

Liu, Shuaiqi, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2022. “Key Phrase Aware Transformer for Abstractive Summarization.” Information Processing & Management 59, no. 3: 102913. doi:10.1016/j.ipm.2022.102913.

Liu, Yang, and Mirella Lapata. 2019. “Text Summarization with Pretrained Encoders.” 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), edited by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, 3730-40. Hong Kong: Association for Computational Linguistics. doi:10.18653/v1/D19-1387.

Liu, Zhengyuan, Angela Ng, Sheldon Lee Shao Guang, Ai Ti Aw, and Nancy F. Chen. 2019. “Topic-Aware Pointer-Generator Networks for Summarizing Spoken Conversations.” In 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 814-21. Singapore: IEEE. doi:10.1109/ASRU46091.2019.9003764.

Martínez-Ávila, Daniel, Fabio Assis Pinho, Francisco Arrais Nascimento, and Richard P. Smiraglia. 2023. “Gay Male Nomenclatures: Bounding the Information Ecology of Desire and Performance in Technologies and Applications.” Knowledge Organization 50: 407-17. doi:10.5771/0943-7444-2023-6-407.

Mendes, Afonso, Shashi Narayan, Sebastião Miranda, Zita Marinho, André F. T. Martins, and Shay B. Cohen. 2019. “Jointly Extracting and Compressing Documents with Summary State Representations.” 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), edited by Jill Burstein, Christy Doran, and Thamar Solorio, 3955-66. Minneapolis, MN: Association for Computational Linguistics. doi:10.18653/v1/N19-1397.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. “Language Models Are Unsupervised Multitask Learners.” OpenAI technical report.

See, Abigail, 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), edited by Regina Barzilay and Min-Yen Kan, 1073-83. Vancouver: Association for Computational Linguistics. doi:10.18653/v1/P17-1099.

Sotudeh, Sajad, Hanieh Deilamsalehy, Franck Dernoncourt, and Nazli Goharian. 2021. “TLDR9+: A Large Scale Resource for Extreme Summarization of Social Media Posts.” In Proceedings of the Third Workshop on New Frontiers in Summarization, edited by Giuseppe Carenini, Jackie Chi Kit Cheung, Yue Dong, Fei Liu, and Lu Wang, 142-51. Online and Dominican Republic: Association for Computational Linguistics. doi:10.18653/v1/2021.newsum-1.15.

Tripathi, Satvik, Dana Alkhulaifat, and Tessa S. Cook. 2026. “Positive Act of Reporting Negative Results in Large Language Model Research: A Call for Transparency.” Journal of the American Medical Informatics Association 33: 926-29. doi:10.1093/jamia/ocaf221.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” In Advances in Neural Information Processing Systems 30, edited by Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, 5998-6008.

Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. “PEGASUS: Pre-training with Extracted Gap-Sentences for Abstractive Summarization.” In Proceedings of the 37th International Conference on Machine Learning, edited by Hal Daumé III and Aarti Singh, 11328-39. Proceedings of Machine Learning Research 119. PMLR.

Zhang, Mengli, Gang Zhou, Wanting Yu, and Wenfen Liu. 2021. “FAR-ASS: Fact-Aware Reinforced Abstractive Sentence Summarization.” Information Processing & Management 58, no. 3: 102478. doi:10.1016/j.ipm.2020.102478.

Zhang, Mengli, Gang Zhou, Ningbo Huang, Peng He, Wanting Yu, and Wenfen Liu. 2023. “AsU-OSum: Aspect-Augmented Unsupervised Opinion Summarization.” Information Processing & Management 60, no. 1: 103138. doi:10.1016/j.ipm.2022.103138.