Domain-specific Guided Summarization for Mental Health Posts

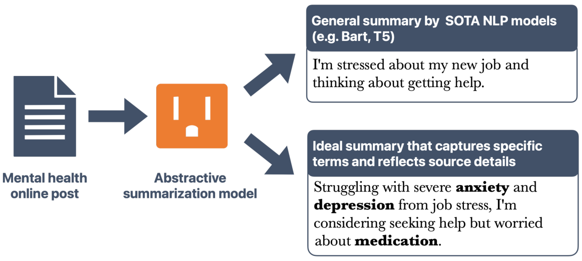
Lu Qian[1,2], Yuqi Wang[1,2], Zimu Wang[1,2], Haiyang Zhang[1],
Wei Wang[1,*], Ting Yu[3], Anh Nguyen[2]
[1]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, China
[2]Department of Computer Science, University of Liverpool, UK
[3]School of Information Science and Technology, Hangzhou Normal University, China
{Lu.Qian21,Yuqi.Wang17,Zimu.Wang19}@student.xjtlu.edu.cn
{Haiyang.Zhang,Wei.Wang03}@xjtlu.edu.cn, yut@hznu.edu.cn, Anh.Nguyen@liverpool.ac.uk

Abstract

In domain-specific contexts, particularly mental health, abstractive summarization requires advanced techniques adept at handling specialized content to generate domain-relevant and faithful summaries. In response to this, we introduce a guided summarizer equipped with a dual-encoder and an adapted decoder that utilizes novel domain-specific guidance signals, i.e., mental health terminologies and contextually rich sentences from the source document, to enhance its capacity to align closely with the content and context of guidance, thereby generating a domain-relevant summary. Additionally, we present a post-editing correction model to rectify errors in the generated summary, thus enhancing its consistency with the original content in detail. Evaluation on the MENTSUM dataset reveals that our model outperforms existing baseline models in terms of both ROUGE and FactCC scores. Although our experiments are specifically designed for mental health posts, the methodology we've developed is intended to offer broad applicability, highlighting its potential versatility and effectiveness in producing high-quality domain-specific summaries.

1 Introduction

Mental health is a critical area that profoundly affects both individuals and society, demanding effective and accurate communication for support (Hua et al., 2024). In this domain, abstractive summarization plays a pivotal role by condensing one lengthy user post from online platforms like Reddit[1] and Reachout[2] into a concise summary. This process, through paraphrasing, generalizing, and reorganizing content with novel phrases and sentences, effectively conveys the essential information and meaning of the original text (Shi et al., 2021; Qian et al., 2023). The summary enables quicker review and response by professional counselors, thus enhancing support for individuals dealing with mental health issues and demonstrating significant social impact.


Mental health online post
Abstractive summarization model
General summary by SOTA NLP models (e.g. Bart, T5)
Ideal summary that captures specific terms and reflects source details

I'm stressed about my new job and thinking about getting help.

Struggling with severe **anxiety** and **depression** from job stress, I'm considering seeking help but worried about **medication**.

Figure 1: This example highlights the importance of an ideal summary that, compared to a general summary, is focused on domain relevance and faithful to the source post, providing essential support for effective communication within the mental health community.

Despite advancements in natural language processing (NLP), applying abstractive summarization to mental health posts illustrates some major challenges in domain-specific summarization. The first challenge is that the summary generated by state-of-the-art (SOTA) pre-trained models (Liu and Lapata, 2019; Lewis et al., 2020; Raffel et al., 2020) tends to be too general and *lacks domain specificity*. These models often struggle to control the content of the summary, making it difficult to determine in advance which parts of the original content should be emphasized (Dou et al., 2021). The second challenge pertains to the *faithfulness* of the generated summary. Often, there is a notable risk of producing a summary that may contradict or diverge from the source document, potentially introducing intrinsic hallucination[3] or inconsistency (Kryscinski et al., 2020; Wang et al., 2024a,b; Na et al., 2024). Together, these issues highlight the need for more advanced summarization techniques that can

[*] Corresponding author.

[1] https://www.reddit.com

[2] https://au.reachout.com

[3] Intrinsic hallucination refers to content in a generated summary that contradicts the source document.

adeptly handle the complexities of domain-specific content while ensuring contextual relevance and detail consistency, as shown in Figure 1.

Drawing inspiration from the GSUM (Dou et al., 2021) framework for its ability to enhance controllability through guidance signal and constrain summary to deviate less from the source document, we introduce a guided summarizer featuring a dual-encoder and an adapted decoder architecture that leverages two types of domain-specific knowledge-based guidance, i.e., specialized mental health terminologies and contextually rich sentences from source post. This design is specifically tailored to enhance the summarization process within mental health contexts, guiding the generation of a summary that is both terminologically precise and richly informed by the underlying domain-specific information contained within the original text.

Further, building on established post-editing practice in recent studies (Dong et al., 2020; Cao et al., 2020), we propose a corrector that follows the summarizer and is dedicated to identifying and correcting potential inconsistencies in the generated summary with respect to the source post. This step ensures the corrected summary more faithfully represents the details of the original text. At last, we evaluate our model on MENTSUM (Sotudeh et al., 2022b), the first mental health summarization dataset. The output summary is evaluated by not only the ROUGE scores (Lin, 2004) measuring linguistic quality, but also FactCC score (Kryscinski et al., 2020), an automatic metric assessing factual consistency[4] with the source document.

The contributions of this study are as follows:

- We introduce novel domain-specific guidance signals, encoded by a separate encoder to guide the summarization process to align closely with the content and context of guidance, thus improving the summary's domain relevance.
- We propose a correction model as a subsequent enhancement step to identify and rectify any potential inconsistency in the generated summary, thereby reducing intrinsic hallucination and further improving faithfulness.
- Our top-performing model, using contextually rich sentences as guidance, outperforms the previous SOTA model CURRSUM (Sotudeh et al., 2022a), achieving improvements of 0.40, 0.82, and 4.07 in ROUGE-1, ROUGE-2, and ROUGE-L scores, respectively. Furthermore, it achieves a 2.5% higher FactCC score compared to BART, and a 3.0% increase over the original GSUM.

2 Related Work

2.1 Guided Abstractive Summarization

The development of neural abstractive summarization has seen significant advancements through the implementation of sequence-to-sequence (seq2seq) framework (Chopra et al., 2016; Nallapati et al., 2016) and the Transformer architecture (Vaswani et al., 2017; Lewis et al., 2020; Raffel et al., 2020). Building on these foundations, guided abstractive summarization leverages additional guidance signals or user input to steer the summarization process, ensuring that the resulting summary is aligned with the specific need and preference.

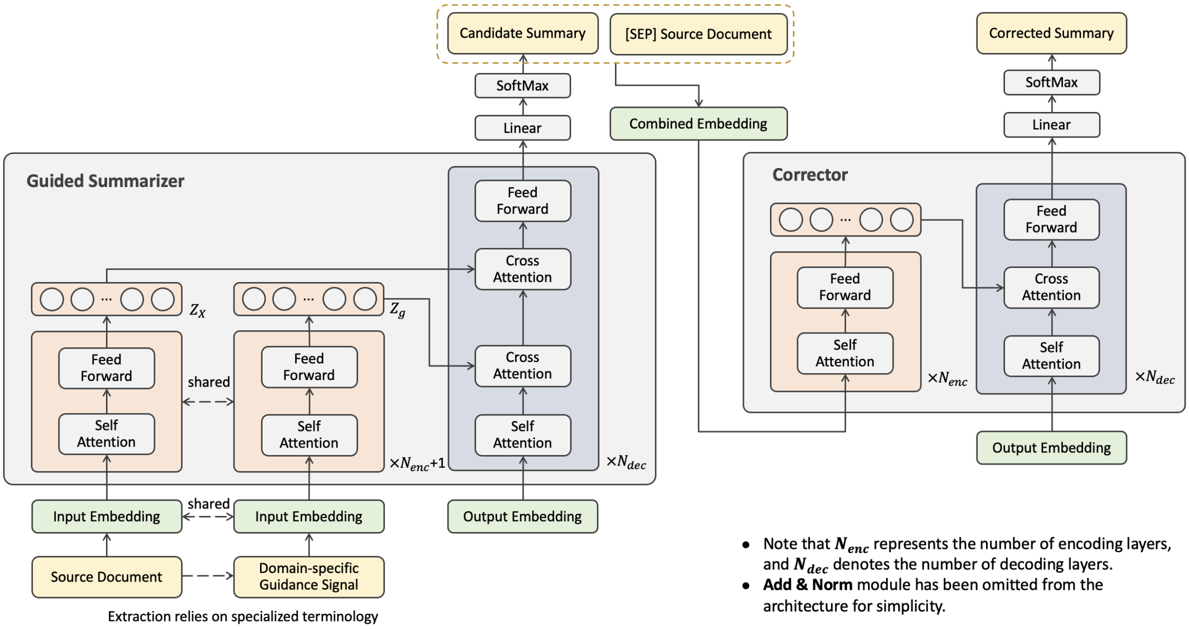
Knowledge bases (KBs) are the most popular guidance and enable summarization systems to deeply engage with the semantic relationship and hierarchical structure they encapsulate. Internal KBs (Huang et al., 2020; Zhu et al., 2021) extract knowledge directly from source documents using information extraction tools (Wang et al., 2024c), reducing intrinsic hallucination and improving the summary's faithfulness. Meanwhile, external KBs (Liu et al., 2021; Dong et al., 2022; Zhu et al., 2024) provide common-sense or world knowledge, enhancing the factuality and reliability of the generated summary.

For other guidance, He et al. (2022) and Narayan et al. (2021) incorporate user-defined keywords and learned entity prompts, respectively. Moreover, Dou et al. (2021) expands on these ideas with the GSUM framework, which supports different types of guidance signals, i.e., highlighted sentences, keywords, salient relational triples, and retrieved summaries.

2.2 Domain-specific Summarization

Domain-specific summarization, particularly in the healthcare field, faces challenges due to the complexity of terminology, the critical need for accuracy in health-related decisions, and the concern over patient confidentiality and data privacy. However, the emergence of advanced NLP techniques and the availability of large annotated med-

[4] Although recent studies define “factuality” as being based on real-world facts, our paper uses the term “factual consistency”, which is commonly employed in evaluation research, to emphasize alignment with the source document.



- Note that N_{enc} represents the number of encoding layers, and N_{dec} denotes the number of decoding layers.
- **Add & Norm** module has been omitted from the architecture for simplicity.

Figure 2: The overall architecture: The initial phase involves a guided summarizer with a dual-encoder and an adapted decoder architecture, utilizing domain-specific guidance signals to produce a candidate summary. This is then refined in the second phase by a post-editing corrector, which identifies and corrects potential inconsistencies in the candidate summary with respect to the source document.

ical datasets have spurred increased interest and progress in this area.

Key efforts include the development of automated radiology report summarization to help streamline healthcare by turning complex radiographic findings into concise summaries, supported by datasets like Indiana University chest X-ray collection (OpenI) (Demner-Fushman et al., 2015) and MIMIC-CXR (Johnson et al., 2019). Similarly, innovative approaches like the Re^{3}Writer model (Liu et al., 2022) leverages the “Patient Instruction” (PI) dataset from MIMIC-III to generate discharge instructions tailored to individual patient records by simulating the physician decision-making process. Additionally, efforts to summarize varied hospital course notes into Brief Hospital Course (BHC) summaries (Searle et al., 2023) utilize adapted BART model, enhanced with clinical ontology signals for producing problem-list-orientated summaries. Furthermore, the creation of the MENTSUM (Sotudeh et al., 2022b) dataset for mental health online posts summarization on Reddit further exemplifies the domain’s growing research interest, with models like CURRSUM employing curriculum learning strategy to improve performance. These advancements highlight the evolving landscape of healthcare summarization, driven by a blend of the latest NLP technologies and domain-specific knowledge.

3 Methodology

The overall architecture of our proposed model is illustrated in Figure 2. By leveraging the strength of both guided summarization and correction in a unified framework, this integrated approach aims to generate summaries that are both domain-relevant and faithful, addressing the challenges of domain-specific summarization.

3.1 Guided Summarizer

Domain-specific Guidance Signal. The core innovation of our model lies in introducing domain-specific guidance signals, encoded by a separate encoder and designed to steer the summarization process to closely align with the content and context of guidance. Specifically, we extract two types of guidance signals from source posts: specialized mental health terminologies and, separately, sentences that contain any of these identified terms. Intuitively, incorporating this knowledge-based guidance would help the summary enhance domain specificity by adhering to specialized terminologies and emphasizing relevant underlying information within the original text (Wang et al., 2023). More details about the guidance extraction are described in Section 4.3.

Dual-encoders. The first encoder transforms source document $X = (x_1, ..., x_n)$ into a sequence of contextual representations $Z_X = (z_{x_1}, ..., z_{x_n})$, while the second encoder processes domain-specific guidance signal $g = (g_1, ..., g_k)$, which can be either terms or sentences, into a sequence of guidance representations $Z_g = (z_{g_1}, ..., z_{g_k})$, where k is the length of the guidance input. Employing self-attention and feed-forward blocks followed by layer normalization, each encoder yields the output vector that encapsulates rich contextual and guidance-driven information for each token in both the document and the guidance.

Decoder. The decoder then integrates outputs from both encoders to generate the summary $Y = (y_1, ..., y_m)$. Modifications have been made to the standard Transformer's decoder structure, enabling it to attend to both the document and the guidance, instead of just one input sequence. Specifically, in each decoding layer, after the self-attention block, the decoder first attends to the guidance representations Z_g, enabling it to decide which part of the source document should be focused on. Then, it uses these signal-aware intermediate representations to more effectively attend to the document representations Z_X, culminating in a summary that is both informative and aligned with the guidance.

Training Objective. The objective function aims to maximize the log-likelihood of generating the summary Y given both the source document X and the guidance signal g. It is formulated as:

$$\begin{aligned} &\arg\max_{\theta} \sum_{i=1}^{N} \log P(Y^{(i)} | X^{(i)}, g^{(i)}; \theta) \\ &= \arg\max_{\theta} \sum_{i=1}^{N} \sum_{t=1}^{m^{(i)}} \log P(y_t^{(i)} | y_{<t}^{(i)}, X^{(i)}, g^{(i)}; \theta), \end{aligned} \tag{1}$$

where N is the number of training examples, $Y^{(i)}$, $X^{(i)}$, and $g^{(i)}$ represent the summary, source document, and guidance for the i-th example, respectively, and θ denotes the learnable parameters of our model. This can be further decomposed into the sum of the log probabilities of each token in the summary conditioned on the preceding tokens, the source document, and the guidance, where $m^{(i)}$ is the length of the i-th summary, and $y_{<t}^{(i)}$ denotes all generated tokens in the i-th summary before position t.

By optimizing this function, our model learns to produce one summary that not only captures the essence of the source document but also closely adheres to the guidance signal. During training, the parameters of the word embedding layers and the bottom encoding layers are shared between the two encoders to reduce the computation and memory requirements, while the top layers of the two encoders are distinct, and initialized with pre-trained parameters but separately trained for each encoder. In the decoder, the first cross-attention block is initialized randomly since it is additional to the standard Transformer structure, while the second cross-attention block is initialized with pre-trained parameters.

3.2 Corrector

In addition to the guided summarizer, we propose a neural corrector as a subsequent enhancement to identify and rectify potential inconsistencies in the generated summary with respect to the source document. This correction process can be modeled as a seq2seq problem: given a candidate summary Y and its corresponding document X, it aims to produce a corrected summary Y' that is more consistent with the original document X.

Artificial Corruption Data. To adequately train the neural corrector, we generate synthetic examples by introducing intentional errors based on heuristics by Kryscinski et al. (2020). This involves creating incorrect summaries by swapping entities, numbers, dates, or pronouns using a strategy outlined by Cao et al. (2020). Specifically, the first three swaps are made by replacing one item in the reference summary with another random item of the same type from the source document, while the pronoun swap is made by replacing one pronoun with another one of a matching syntactic case.

Model Design. The correction model is designed to rectify an incorrect summary Y into a consistent summary Y' with minimal modifications based on the source document X. This can be formulated as optimizing the model parameters θ to maximize the likelihood function within an encoder-decoder framework:

$$\arg\max_{\theta} \sum_{i=1}^{N} \log P(Y'^{(i)} | Y^{(i)}, X^{(i)}; \theta), \tag{2}$$

where N is the number of synthetic training examples, and θ denotes the model parameters.

For this purpose, we use BART (Lewis et al., 2020) as the foundation for fine-tuning the corrector due to its proven effectiveness in conditional text generation tasks. BART is a seq2seq autoregressive transformer pre-trained on various denoising objectives, such as text infilling and token deletion, making it adept at recovering the original text from corrupted input. This pre-training aligns naturally with our summary correction task, where the model treats the incorrect summary as noisy input, focusing on resolving errors to recover factual consistency.

4 Experiments

4.1 Dataset

Our research utilizes MENTSUM, the first mental health summarization dataset, which contains selected user posts from Reddit along with their short user-written summaries (called TL;DR) in English. Each lengthy post articulates a user's mental health problem and quest for support from community and professional counselors, while the corresponding TL;DR serves to condense this narrative into a concise summary, facilitating quicker review and response by counselors. This dataset comprises over 24k post-TL;DR pairs, divided into 21,695 training, 1,209 validation, and 1,215 test instances. On average, each post contains 327.5 words or 16.9 sentences, while TL;DR consists of 43.5 words or 2.6 sentences. More details about the dataset can be found in Sotudeh et al. (2022b).

4.2 Metrics

To evaluate the linguistic quality of the generated summary, we use standard ROUGE metrics: ROUGE-1, ROUGE-2, and ROUGE-L. These metrics assess the overlap of unigrams, bigrams, and the longest common subsequence, respectively, between the generated summary and reference one. We report the F1 scores for these metrics to provide a comprehensive analysis.

For automatically assessing the factual consistency of the generated summary with the source document, we utilize a fine-tuned version of the FactCC model (Kryscinski et al., 2020). This model maps the consistency evaluation as a binary classification problem, and outputs a probability score ranging from 0 to 1, indicating the likelihood that the generated summary is factually consistent with the source content.

4.3 Implementation Details

Guided Summarizer. To construct knowledge-based guidance, we curate mental health terminologies from subsets released by Kaiser Permanente (KP) in 2011 and 2016[5], focusing on the "KP_Patient_Display_Name" column. Our preprocessing involves (1) separating terms that are combined with commas to ensure each term is individually identifiable, (2) splitting terms that contain parentheses (e.g., "A (B)") into two separate entities to simplify and clarify the data, (3) removing duplicates to compile a list of unique terms, and (4) excluding terms longer than three words to improve regex matching efficiency. This process yields a refined list of 1,068 unique terminological terms. Then, we extract these identified terms from each mental health post, separate them with a special [SEP] token, and use them as the first type of guidance. Additionally, we explore an alternative approach by extracting sentences from each source post that contain any of the predefined terminology, using them as the second type of guidance. Regular expressions are employed to ensure a precise match of the entire term, avoiding partial or irrelevant matches.

We adopt the BART-large as the foundation for fine-tuning our guided summarization model[6]. Training parameters include a total of 10,000 updates, a maximum token of 1,024, and an update frequency of 4. We opt for the AdamW optimizer with a learning rate of 3e-5, β parameters set to (0.9, 0.98), and a weight decay of 0.01. The objective function is cross-entropy Loss across all models. After training for five epochs, the model checkpoint achieving the highest ROUGE-L score on the validation set is selected for inference. For decoding, we employ a beam size of 6, with minimum and maximum lengths set to 15 and 200, respectively, and a restriction on repeating trigrams. All our experiments are conducted on four NVIDIA Tesla V100 GPUs, with the training process requiring approximately four hours.

Error Corrector. We create synthetic incorrect summaries incorporating entity, number, date, and pronoun errors, resulting in 25,940 training and 1,416 validation examples. Based on the BART-

[5] https://www.johnsnowlabs.com/marketplace/cmt-mental-health-problem-list-subset/

[6] https://github.com/neulab/guided_summarization

Model	Guidance Signal	Rouge-1	Rouge-2	Rouge-L	100×FactCC
CurrSum	No signal	*30.16*	*8.82*	*21.24*	–
Bart	No signal	28.792	8.741	23.657	87.74
After Correction		28.754	8.722	23.625	88.40 (↑0.75%)
GSum	Highlighted sentences	30.031	8.917	24.698	87.65
After Correction		30.013	8.907	24.685	87.98 (↑0.38%)
GSum-term	Specialized terminologies	30.429	9.441	**25.335**	89.05
After Correction		30.426	9.425	25.326	89.22 (↑0.19%)
GSum-sent	Context-rich sentences	**30.578**	**9.647**	25.315	90.12
After Correction		30.561	9.638	25.309	**90.62** (↑0.55%)

Table 1: Rouge scores and FactCC scores on MentSum test set.

large architecture implemented in fairseq[7], the neural corrector is fine-tuned with the parameter setting similar to the guided summarizer, except it is trained for 10 epochs to allow the model to adequately learn to identify and correct these subtle errors. During inference, the candidate summary generated from the previous guided summarizer is concatenated with its source post, and processed by the optimal checkpoint to produce the corrected summary for final evaluation.

FactCC Evaluator. We re-implement and fine-tune the FactCC model[8], tailoring it to better suit our domain-specific needs. The training data consist of both correct and incorrect examples: the former derives from clean reference summary (labeled as "CORRECT"), while the latter uses the same synthetic data as the corrector (labeled as "INCORRECT"), signifying inconsistent with the source post. Thus, we obtain 21,695 correct and 25,940 incorrect examples for training, with 1,209 correct and 1,416 incorrect examples for validation. Based on the Bert-base model, we use the same hyper-parameters for training the original FactCC model over 10 epochs. For inference, the corrected summary (defined as "claim") and its corresponding source post (defined as "text") are combined and fed into the optimally selected checkpoint (with the lowest Loss) to compute a probability score, quantitatively evaluating the alignment between claim and text.

4.4 Baselines

Bart. It is a pre-trained SOTA model for summarization tasks, and demonstrated superior performance over various extractive and abstractive summarizers on MentSum dataset (Sotudeh et al., 2022b). We re-employ Bart on this dataset as a baseline rather than simply copying the results because that study did not evaluate factual consistency, a key focus of our research for comparison. In this baseline experiment, training parameters match those of the guided summarizer, with the exception of setting the update frequency to 1.

GSum. We adopt GSum with highlighted sentences, the best-performing guidance signal, as our second baseline. Highlighted sentences are identified as oracle sentences during training using a greedy search algorithm for maximum Rouge scores with reference summaries, but are extracted during inference by employing a pre-trained extractive summarizer, i.e., the best-performing BertExt checkpoint (Liu and Lapata, 2019), due to unseen references summaries in the test set. This baseline experiment proves more complex compared to our guided summarizer, as it requires an additional summarizer during inference, a limitation within the original framework. In contrast, our guidance extraction (described in Section 4.3) relies on specialized terminologies, moving away from reliance on oracle sentences, and streamlining the process by unifying the training and inference phases.

5 Results

5.1 Automatic Evaluation

We refer to the model that utilizes extracted mental health terminologies for guidance as GSum-term, and the model that employs context-rich sentences containing these domain-specific terms for guidance as GSum-sent. As shown in Table 1, experimental results on the MentSum dataset reveal that our proposed models surpass established benchmarks and baseline models in terms of both Rouge

[7] https://github.com/pytorch/fairseq/blob/master/examples/bart

[8] https://github.com/salesforce/factCC

User posts	Human-written	GSUM-SENT generated	Corrected
... *About 6- 8 months later ther **anxiety** was back* ... So back to the prozac I went. I took it again for about 18 months and then the cycle continued ... *I finally went to see a psychiatrist and she suggested I go back on it again since my **anxiety** has gotten out of control again* ...	going back on Prozac for the third round ...	I started taking prozac about 4 years ago, stopped taking it after 18 months and then took it again for about 6 - 8 months. Now I 'm back on it again and my **anxiety** has gotten out of control again ...	... then took it again for about 18 months ...
*I'm 19 year old male with **autism, depression and anxiety*** ... *Since then my dad has sunk into a deep **depression***... and he has threatened to kill himself ... My mother who is also extremely sickly has dealt with him for years ... and is probably going to leave him. On top of all this we cannot afford to continue to live in our house ...	mom and dad sick, dad threading suicide, mom leaving dad, we re losing our house ...	I m a 19 year old male with **autism, depression and anxiety** who ca nt afford to live in our house, my father is **depressed** and suicidal, my mother is leaving me ...	... my mother is leaving him ...
... made an excellent suggestion w CBT, but if cost or availability is an issue, guidance is a great way to start ... *When I was first coming to grips with my **anxiety**, I journaled quite frequently. Not only was it a great **emotional** release to let my intrusive thoughts out* ...	I suggest looking into CBT or a guidance counselor. Journalling is good ...	CBT is a great way to start. Do n't be afraid to journal ...	Guidance is a great way to start ...
... *I've been really **depressed** for years* ... I'll spare you the details but this goes on for close to 10 years ... *I've been failing everything I take due to a combination of **depression, anxiety*** ...	My life was messed up ...	I 've been in and out of mental institutions for 10 years and have been failing everything I take due to a combination of **depression, anxiety** ...	I 've been in and out of mental institutions for years ...

Table 2: Four samples of mental health posts along with their human-written TL;DRs, summaries generated by GSUM-SENT, and corrections applied by neural corrector.

and FactCC scores. Below, we present more detailed insights derived from our experiments.

Effectiveness of Guided Summarizer. Initial observations from two baseline experiments indicate that guided summarizer exhibits improved ROUGE scores, particularly in the ROUGE-2 and ROUGE-L metrics, compared to CURRSUM, indicating a better capture of detailed information and narrative structure. However, the original GSUM achieves a lower FactCC score compared to BART, suggesting that while highlighted sentences can steer the model toward relevant information, they do not guarantee factual consistency.

Improvement through Domain-specific Guidance. Our experiments with the proposed models yielded significant improvements on both ROUGE and FactCC scores over the baseline models, indicating improvements in summary quality and factual consistency. Specifically, GSUM-TERM is 1.5% higher than BART and 1.6% higher than GSUM on FactCC score, suggesting that the use of specialized terminologies as guidance signal, instead of highlighted sentences, is effective in enhancing the summary's alignment with the source content while maintaining or even improving its overall quality.

The subsequent experiment with the GSUM-SENT model employs context-rich sentences embedded with domain-specific terms as the guidance signal, leading to notable advancement across the board. Specifically, the model not only records superior ROUGE scores but also achieves a 2.7% higher FactCC score compared to BART and 2.8% improvement over GSUM. This finding, resonating with the insight from the original GSUM study, highlights the superiority of contextually rich, sentence-based guidance over simpler keyword-based one. Overall, this integration of domain-specific guidance underscores the importance of leveraging specialized information from the source post, and is pivotal for the generated summary to improve its alignment with the source content in the mental health context.

Benefit of Corrector. The correction model demonstrates its capability to refine the consistency of summary and faithfully represent the source details across both our proposed models and baseline models. After correction, the FactCC scores showed absolute improvements ranging from 0.17 to 0.66 percentage points and relative increases between 0.19% and 0.75% across all evaluated models. It's worth noticing that correction generally results in a slight decline in ROUGE scores, a phenomenon observed in multiple studies (Kryscinski et al., 2020; Maynez et al., 2020), and may be attributed to the nuanced balance between enhancing factual consistency and maintaining linguistic quality in the summary.

5.2 Case Study and Analysis

Acknowledging the limitations of automatic evaluation in the summarization system, we also manually assess the quality of our work by comparing candidate summaries generated by GSUM-SENT and corrected ones against human-written TL;DRs, as shown in Table 2. To protect user privacy, the source posts are selectively displayed. The specialized mental health terminologies are highlighted in **bold**, and sentences containing these terms are in *italic* to show their influence on the summary generation process. Additionally, corrections and related text segments are marked in red to provide clear insight into the improvements in detail consistency.

Heightened Domain Specificity. Summaries generated by GSUM-SENT often capture more specialized mental health terms. Conversely, TL;DRs are written in a colloquial and condensed manner, which might omit essential terminological details. Taking the first sample as an example, the human-written summary merely mentions going back on Prozac for the third time, while the GSUM-SENT-generated one specifies details on the duration of treatment and the underlying issue of anxiety. Similarly, in the fourth sample, the human-written summary describes the situation as “messed up”, a vague term compared to the explicit mentions of “depression” and “anxiety” by GSUM-SENT. They all indicate the model’s potential to provide more transparent communication of mental health issues, which is helpful when asking for support from professional counselors.

Improved Faithfulness. Both the guided summarizer and corrector play crucial roles in improving faithfulness according to reported FactCC scores, with the corrector further enhancing detail consistency with respect to the source post. It addresses date inaccuracy in the first and fourth samples, corrects pronoun usage in the second, and resolves entity error in the third. These errors originate from incorrect references to similar items within the original posts, exemplified by the misrepresentation of “6-8 months” in the first sample.

Despite the precision in correction, there is a shortcoming: the corrector’s modifications are very subtle, attributed to its training on a dataset limited to four types of minor errors. This restraint in correction is evident in our examination of summaries generated by GSUM-SENT model, where only 10.3% undergo revisions by corrector. Moreover, these adjustments are minimal, with 92.8% of the corrected summaries incorporating three or fewer new tokens, despite the summary averaging 53.27 tokens in length. This indicates that the current correction model may not fully capture complex inaccuracies beyond its training scope, highlighting the need for a more diverse training dataset to enhance its ability to improve detail consistency across a wider range of summaries.

6 Conclusion

Focusing on the mental health domain, our research addresses the challenges of generating domain-relevant and faithful summaries through the development of a guided summarizer followed by a neural corrector. By incorporating novel domain-specific knowledge-based guidance, especially context-rich sentences, our adapted summarizer closely aligns with the specialized source content and effectively enhances the domain relevance of the generated summary. The post-editing corrector further ensures the elimination of inconsistency or intrinsic hallucination, making the summary more faithful to the source document.

Comprehensive evaluation with the MENTSUM dataset demonstrates the superior performance of our proposed model over existing baselines, as evidenced by improvements in both ROUGE and FactCC scores. Although our experiments are specifically tailored to the mental health domain, the methodologies we’ve developed are designed to be adaptable across various fields where the precision of domain-specific knowledge and detail consistency are both essential, such as in legal, financial, or technical contexts. The demonstrated effectiveness and adaptability of our approach underscore its potential to advance domain-specific abstractive summarization, offering a versatile framework for future exploration.

Acknowledgments

We thank to anonymous reviewers for their valuable comments, and Georgetown University for providing the MENTSUM dataset. This work is partially supported by the 2022 Jiangsu Science and Technology Program (General Program), contract number BK20221260.

References

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331, Online. Association for Computational Linguistics.

Yue Dong, John Wieting, and Pat Verga. 2022. Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1067–1082, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zi-Yi Dou, 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*, pages 4830–4842, Online. Association for Computational Linguistics.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi han Sheu, Peilin Zhou, Lauren V. Moran, Sophia Ananiadou, Andrew Beam, and John Torous. 2024. Large language models in mental health care: a scoping review. *Preprint*, arXiv:2401.02984.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5094–5107, Online. Association for Computational Linguistics.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1).

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. 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, Online. 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.

Fenglin Liu, Bang Yang, Chenyu You, Xian Wu, Shen Ge, Zhangdaihong Liu, Xu Sun, Yang Yang, and David Clifton. 2022. Retrieve, reason, and refine: Generating accurate and faithful patient instructions. In *Advances in Neural Information Processing Systems*, volume 35, pages 18864–18877. Curran Associates, Inc.

Yang Liu 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)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S. Yu. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6418–6425.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Mieradilijiang Maimaiti, Tong Chen, Wei Wang, Tao Shen, and Ling Chen. 2024. Rethinking human-like translation strategy: Integrating drift-diffusion model with large language models for machine translation. *Preprint*, arXiv:2402.10699.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Lu Qian, Haiyang Zhang, Wei Wang, Dawei Liu, and Xin Huang. 2023. Neural abstractive summarization: A brief survey. In *2023 IEEE 3rd International Conference on Computer Communication and Artificial Intelligence (CCAI)*, pages 50–58.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Thomas Searle, Zina Ibrahim, James Teo, and Richard J.B. Dobson. 2023. Discharge summary hospital course summarisation of in patient electronic health record text with clinical concept guided deep pre-trained transformer models. *Journal of Biomedical Informatics*, 141:104358.

Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K. Reddy. 2021. Neural abstractive text summarization with sequence-to-sequence models. *ACM/IMS Transactions on Data Science*, 2(1):1–37.

Sajad Sotudeh, Nazli Goharian, Hanieh Deilamsalehy, and Franck Dernoncourt. 2022a. Curriculum-guided abstractive summarization for mental health online posts. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 148–153, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Sajad Sotudeh, Nazli Goharian, and Zachary Young. 2022b. MentSum: A resource for exploring summarization of mental health online posts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2682–2692, Marseille, France. European Language Resources Association.

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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yuqi Wang, Wei Wang, Qi Chen, Kaizhu Huang, Anh Nguyen, Suparna De, and Amir Hussain. 2023. Fusing external knowledge resources for natural language understanding techniques: A survey. *Information Fusion*, 92:190–204.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024a. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 470–475, Bangkok, Thailand. Association for Computational Linguistics.

Zimu Wang, Wei Wang, Qi Chen, Qiufeng Wang, and Anh Nguyen. 2024b. Generating valid and natural adversarial examples with large language models. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1716–1721.

Zimu Wang, Lei Xia, Wei Wang, and Xinya Du. 2024c. Document-level causal relation extraction with knowledge-guided binary question answering. *Preprint*, arXiv:2410.04752.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 718–733, Online. Association for Computational Linguistics.

Fangwei Zhu, Peiyi Wang, and Zhifang Sui. 2024. Reducing hallucinations in entity abstract summarization with facts-template decomposition. *arXiv preprint arXiv:2402.18873*.