Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/131/2024.20546

Survey of Context-Sensitive Processing on Dialogue Generation Model

Qichen Zhang[1,3,*,†], Zhiting Lei[2,4,†]

[1]Artificial Intelligence, Beijing Normal University·Hong Kong Baptist University United International College, Zhuhai, 519087, China
[2]Queen Mary University of Hainan, Beijing University of Posts and Telecommunications, LingShui, 100876, China

[3]t330034072@mail.uic.edu.cn
[4]leizhiting2023@gmail.com
*corresponding author
[†]These authors contributed equally and shoule be considered co-first authors.

Abstract. In recent years, dialogue generation models have emerged as a prominent area of research, as a crucial role in the domain of natural language processing, garner significant attention from the academic community. Existing dialogue generation models predominantly emphasize human-computer interaction, however, various context-sensitive issues should not be overlooked inherent to the process. Consequently, this paper aims to summarize and categorize these context-sensitive processing issues. Firstly, based on extant literature, an overview of the current landscape of dialogue generation models is provided. Secondly, the definition of "sensitive" is clarified, and relevant scholarly works are briefly reviewed following the definition. Thirdly, previous methods addressing context-sensitive processing within dialogue generation models are directly categorized and their specific methodologies are briefly analyzed regarding this issue. Finally, pertinent research gaps and limitations are identified while future directions for research in dialogue generation models are proposed.

Keywords: Human-machine dialogue, Machine learning, Dialogue generation model, Sensitive Content Processing.

1. Introduction

Research on human-computer dialogue systems has been ongoing for many years and continuously iterating—from the earliest proposal of the Turing Test to the recent advent of GPT, the study of human-computer interaction has always been a focal point for both academia and industry. Existing research on dialogue systems can be categorized into task - oriented dialogues and non-task-oriented dialogues[1]. The former, with the development of deep learning, possesses the ability to learn data features and minimize human interference, while the latter is primarily constructed through information retrieval and generative methods, ensuring that responses are coherent in theme, diverse in language, and personalized. Notably, generative methods take greater account of contextual influences, better simulating normal human communication[1]. In recent years, with the emergence of large corpora, dialogue generation models have become a hot topic in academic research. However, the process of exchanging and

DOI: 10.54254/2755-2721/131/2024.20546

communicating information between humans and machines inevitably involves the confusion of sensitive information. Sensitive information is varied and complex; without proper handling, ethical, political, and religious implications can arise during human - machine interactions. Currently, most studies focusing on dialogue generation models prioritize emotion-aware dialogue systems[2], but sensitive issues extend beyond emotional sensitivity. This paper expands the scope of text sensitivity processing and elaborates on its classifications.

2. Related Work

Since most literature focuses on emotional dialogue generation, the related work section primarily outlines studies based on emotion-sensitive topics and extracts their context -sensitive processing content. Additionally, as many papers do not emphasize context-sensitive processing, existing reviews have not provided timely and comprehensive summaries of the latest research in this area.

Gieselmann and Ostendorf[3] addresses error handling and problem detection in human-robot dialogues in 2007, focusing on improving response strategies by identifying misunderstandings and help requests. This approach offers valuable insights for managing sensitive topics, suggesting that similar error detection mechanisms could be applied to identify potentially sensitive or inappropriate subjects. Nowadays, Xue et al.[4] presents E-chat, an emotion-sensitive dialogue system leveraging large language models, designed to generate appropriate responses based on the emotional cues in users speech in 2023. This introduces a novel strategy for managing sensitive content, suggesting that emotion detection can assist the system in avoiding inappropriate responses when users are in a negative emotional state.

At the same time, emotional sensitivity in model feedback is also an important direction, aimed at guiding users' positive emotions. The paper "Eliciting Positive Emotion through Affect- Sensitive Dialogue Response Generation: A Neural Network Approach"[5], published in 2018, examines how emotion-sensitive generation mechanisms can evoke positive emotions in users, demonstrating that neural networks, through emotion encoders, can ensure emotional consistency throughout dialogues. This has significant implications for handling sensitive topics, as emotion-sensitive response generation can help prevent the provocation of negative emotions. In addition, the model can use emotion tracking to controlling the handling of sensitive topics in dialogue generation , particularly in generating contextually appropriate responses[6].

Not only are there the aforementioned handling of emotional sensitivity issues, but also the management of sensitive issues, which relates to safety and fairness issue of generations from dialogue models, such as personal privacy information. A method for generating context - sensitive dialogues based on the Seq2Seq model, which is sensitive to speaker names , has been introduced by Jia et al.[7].

The aforementioned related literature covers multiple aspects of sensitive text processing. This paper distills the essence of the aforementioned literature and focuses on dialogue generation models as the research subject, summarizing and describing key issues related to context-sensitive processing. It also discusses the creation and application of dialogue generation models in context-sensitive processing. Finally, it highlights the development gaps and future directions for context-sensitive processing in dialogue generation models.

3. Definition

Because the definition of "sensitive" in this paper is relatively broad, it is reiterated here. The definition of sensitive in this paper specifically referring to sensitive words (such as swearwords), sensitive topics (such as extreme left and right political topics), personal privacy topics, and emotionally sensitive topics (such as extreme emotions during dialogue). Therefore, the various processing methods described in the following article will be applied to the above context-sensitive problems separately(or maybe applied simultaneously).

DOI: 10.54254/2755-2721/131/2024.20546

4. Thematic Grouping

Aiming at the pertinent literatures, this paper extracts context-sensitive processing content, classifies and analyzes it.

Discussing how to deal with context-sensitive situations during the generation of generative dialogue models, this paper divides them into two categories: one is processing when the model is created, and the other is processing when the model is applied. This classification is inspired by Task- oriented conversations and Non-task-based dialogue.[1]. At the sametime, we hope to reflect the ethical rigor of human-computer interaction in the construction of the model, as well as the application concept based on the humanistic position.

Therefore, the former (processing when the model is created) mainly involves the design of the model itself–it can be understood as the artificial making the model conform to the necessary ethical requirements, including: the artificial data cleaning in the early stage of machine learning and the artificial evaluation in the later stage to identify sensitive issues, the addition of content recognition filters (keywords/text themes/privacy issues) and soon; The latter (processing when the model is applied) mainly involves the emotion of the dialogue with the user or self-generation, that is, human-computer interaction performing operations based on the user side, including: sensitive emotion recognition (in order to follow or improve the emotional direction of the dialogue), contextual content perception (in order to better understand and generate the dialogue with sensitive emotions) and so on.

5. Methodological Analysis

5.1. Creating Models

In the category of "context-sensitive processing when creating models," we can consider that model creation is supervised and requires evaluation. Therefore, we focus on relevant operations during model design to prevent the generation of sensitive words and avoid multi - turn conversations on sensitive topics.

During the model training process, context-sensitive measures are taken, such as data cleaning and filtering at the early stages of machine learning, removing or replacing data related to sensitive words (such as profanity), before feeding this data into the learning process. This ensures that the model itself does not autonomously generate sensitive words or reduces the likelihood of generate sensitive words. This is an important step in early sensitive word processing.

At the output end of the dialogue generation model, a content filter is set up to detect the conversations generated by the model and shield or replace sensitive texts. The content filter here includes: keyword lists, regular expressions, or other text classification models. These content filters can be configured according to the different needs of the model during the creation process, artificially limiting its related generated content. The setup of the content filter can help the model recognize special topics and filter them, enabling the model to avoid outputting certain highly sensitive topics (such as far-left or far-right political topics) and personal privacy topics during the conversation generation process.

The handling of sensitive topics can be achieved through keyword extraction to determine "whether it is a sensitive topic" and take measures, as well as by guiding or predicting the direction of sensitive topics through contextual association. Regarding the handling of sensitive topics, this can be further divided into topic word extraction and context-sensitive association:

At the beginning , it is necessary to define inappropriate content and sensitive topics. "Inappropriate content" does not necessarily include explicit toxic language or abusive words, but rather opinions or stances that are unsuitable for specific topics. These inappropriate contents and sensitive topics generally include the following: pornography, gambling, drugs, suicide, violent crime, race issues, religion, war, etc.

To detect inappropriate content and sensitive topics, pre-trained natural language processing models, such as the BERT model, can be used for deep learning-based detection on specialized datasets, as well as generating negative samples based on specific features to help the model better identify potential sensitive or inappropriate content. Combining text detection with visual feature-based detection, audio

DOI: 10.54254/2755-2721/131/2024.20546

and multimodal analysis detection, and deep learning-based detection can better meet the needs of complex sensitive content detection[6].

About context-sensitive association, Ma et al.[8], Sordoni et al.[9], and Ling et al.[10] have proposed different approaches. Ma and colleagues addressed the issue that the Sequence to Sequence (Seq2Seq) model lacks presentation of sentimental information and dialogue context in the encoding procedure, resulting in generating responses that are poor in sentiment and irrelevant to the context. They proposed a solution: a new sentimental and context - sensitive Seq2Seq model for dialogue generation. This involves using a sentimental word vector and an utterance sentiment vector to reinforce the presentation of sentimental information for the Seq2Seq model, ensuring it contains sentiments. Moreover, this model employs a pronunciation encoder and a context encoder to ensure strong contextual relevance during generation, even allowing for more emotionally charged responses[8]. Sordoni et al. addressed two context-sensitive models based on the architecture of the Recurrent Neural Network Language Model (RLM), both encoding past information into a hidden continuous representation and using RLM to decode it, ensuring reasonable contextual associations[9]. The distinctive feature of this model compared to other context-sensitive models is that it does not require manual annotation and can be trained end-to-end on a large amount of social media data. Meanwhile, Ling et al. proposed the Context-Controlled Topic-Aware neural response generation model, also known as CCTA, which is the most relevant research model for handling sensitive topics with a context-sensitive approach in this paper. Its characteristic lies in capturing the semantic relationships and information within a topic, especially at semantic transitions, allowing for a better understanding of the topic within the context[10].

This series of model constructions are effective methods for handling sensitive texts.

5.2. Applying Models

In the context of "context-sensitive processing when models are applied", we focus on how dialogue generation models address sensitive emotions during human-computer interactions. Sensitive emotions in such dialogues refer to emotional states that may trigger fluctuations, misunderstandings, or negative reactions in users. These emotions are often linked to psychological vulnerabilities, making users more susceptible to external triggers, such as responses generated by the system. If not handled properly, these emotions can lead to a negative user experience, task failure, or even a loss of trust in the system. Our review of the relevant literature reveals that sensitive emotions typically manifest as heightened emotional responses, such as anger, sadness, fear, or anxiety. These emotions are often tied to moments of psychological vulnerability, particularly when users face challenges, failures, or distressing situations. Sensitive emotions are also frequently connected to sensitive topics, including death, health issues, personal failures, and socially controversial subjects. When users engage in discussions around these topics, emotional instability is often observed[11].

To scientifically handle users' sensitive emotions during human-computer interactions, the first critical step is enabling the model to accurately perceive user emotions. Numerous studies highlight the importance of emotion detection in dialogue systems, especially when managing sensitive topics. Accurate emotional state detection helps prevent the generation of inappropriate or emotionally triggering content. In this regard, Xue et al.[4] have developed a novel voice dialogue system called E-chat, which integrates RainTone emotion embeddings with large language models. The system employs the HuBERT model as a speech encoder to extract emotional and acoustic features from speech. These features are then processed by a Transformer-based linking module, which converts them into a format compatible with LLM decoders, ensuring the generated text is consistent with the user's speech. Additionally, a "problem indicator" is used to detect sensitive emotions or topics within the dialogue, allowing the system to identify and respond to these issues based on logical rules in a timely and flexible manner[3].

Building on these advancements, Wang et al.[11] have further explored the dynamic changes in user emotions during dialogues and proposed a method called SEEK(Serial Encoding and Emotion-Knowledge Interaction). This approach introduces a fine-grained emotion encoding strategy to

DOI: 10.54254/2755-2721/131/2024.20546

sensitively capture emotional shifts in conversations and fosters an interaction between emotional states and commonsense knowledge, thereby enhancing the quality of empathetic dialogue generation.

After identifying sensitive emotions in a conversation, the next focus of research shifts to "how to respond to these emotions". Gieselmann et al.[3] have proposed a "four-state finite state machine" to manage dialogue state transitions, allowing the system to switch to specific states based on the user's emotional signals and maintain the flow of the conversation. For instance, when the system detects signs of unease or anger in the user's tone, it can enter a "help state" to generate more empathetic responses aimed at relieving the user's stress or confusion. Similarly, a "problem-sensitive response generation mechanism" adjusts the level of care in response to fluctuations in the user's emotions.

Further studies have explored how dialogue systems can guide users toward more positive emotional states, especially when addressing sensitive topics. This strategy holds significant potential for mitigating the negative impact of sensitive subjects in conversations. One such approach is an extension of the Hierarchical Recurrent Neural Network (HRED) architecture, called Emo-HRED, which incorporates an emotion encoder into the traditional HRED model[5]. By capturing and considering the user's emotional state during response generation, Emo-HRED can foster more positive emotional engagement.

6. Conclusion

Discussing context-sensitive processing, we find that the real-time sensitivity handling of existing models during user interactions still needs improvement.

6.1. Localization and personalization features

Current models focus primarily on identifying the main topics in text to retrieve relevant information for generating topic-aligned responses, often overlooking ethical and moral considerations within "topic identification". As a result, when confronted with multi-turn dialogues on sensitive subjects, few models will actively refuse to continue such discussions. In addition, Existing models still face challenges in addressing diverse users, as they often lack sufficient adaptability to different user backgrounds (such as culture, gender, and age). When engaging with such user groups, these models may fail to correctly identify and appropriately respond to sensitive topics. This leads to dialogue generation models lacking localization and personalization features unless an additional agent is introduced.

6.2. Localization and personalization features

Many studies emphasize the critical role of emotions in dialogue–misjudging emotions may result in content that mismatches the user’s emotional state, potentially leading to negative reactions. However, achieving precise emotion detection remains a significant challenge. We expect that dialogue generation models, through user interaction, will be able to predict emotional trends, respond promptly to sensitive emotions, and offer positive guidance where appropriate[12].

According to the predictive model proposed by Chang and Danescu-Niculescu-Mizil in 2019[13], dialogue deviation (in this predictive model, it specifically refers to deviations towards conversations labeled for antisocial events) is highlighted. It indicates that a conversation is heading towards derailment before it actually turns toxic and provides a relatively comprehensive predictive model. This model can be applied to context-sensitive processing, where a detection node is set up at "a conversation is heading towards derailment before it actually turns toxic", allowing the dialogue generation model to stop outputting and provide silent feedback or positive guidance when such nodes are detected.

In summary, we hope that the dialogue generation models can also perform well in practical applications, ensuring more efficient and intelligent processing during human-machine interactions, maintaining basic ethical and moral principles between humans and machines, and improving user satisfaction when using dialogue generation models[14].

DOI: 10.54254/2755-2721/131/2024.20546

7. Conclusion

With the development of large language model technology, more and more researchers have paid attention to dialogue generation model, and it is also the direction of researchers to construct dialogue with low text sensitivity and conform to human cognition. Therefore, the text-sensitive problem processing of dialogue generation model is concerned and developed by many models. In this paper, the text sensitive processing of dialogue generation model is reviewed. Firstly, the existing researches are sorted out and summarized, and the solutions are summarized. Secondly, it summarizes the dialogue model methodology needed in the present research. Finally, some limitations of the existing basic model are proposed, and the description of the full text is summarized.

Acknowledgments

Qichen Zhang and Zhiting Lei contributed equally to this work and should be considered co-first authors.

References

[1] WANG Chunyu, MA Zhiqiang, DU Baoxiang, JIA Wenchao, WANG Hongbin, BAO Caijilahu. Survey of Research on End-to-End Emotional Dialogue Generation[J]. Journal of Frontiers of Computer Science and Technology, 2022, 16(2): 280-295.

[2] Yin, Z., Zhen, L., Tingting, L., Yuanyi, W., Cuijuan, L., & Yanjie, C. (2021). Survey of Affective-Based Dialogue System.

[3] Gieselmann, P., & Ostendorf, M. (2007, September). Problem-sensitive response generation in human-robot Dialogs. In Proceedings of the 8th SIGDial Workshop on Discourse and Dialogue (pp. 219-222).

[4] Xue, H., Liang, Y., Mu, B., Zhang, S., Chen, Q., & Xie, L. (2023). E-chat: Emotion-sensitive Spoken Dialogue System with Large Language Models. arXiv preprint arXiv:2401.00475.

[5] Lubis, N., Sakti, S., Yoshino, K., & Nakamura, S. (2018, April). Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32,No. 1).

[6] Povedano Álvarez, D., Sandoval Orozco,A. L., García-Miguel, J. P., & García Villalba, L. J. (2023). Learning strategies for sensitive content detection. Electronics, 12(11), 2496.

[7] Jia,Q., Tang, H., & Zhu, K.Q. (2023). Reducing Sensitivity on Speaker Names for Text Generation from Dialogues. ArXiv, abs/2305.13833

[8] Ma, Zhiqiang; Du, Baoxiang; Shen, Ji; Wang, Chunyu; Yang, Rui. A Sentimental and Context-Sensitive Model for the Seq2Seq- Based Dialogue Generation. Elektrotehniski Vestnik; Ljubljana Vol. 87, Iss. 3, (2020): 127-134.

[9] Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J., Gao, J., & Dolan, W.B. (2015). A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. North American Chapter of the Association for Computational Linguistics.

[10] Ling, Y., Cai, F., Hu, X., Liu, J., Chen, W., & Chen, H. (2021). Context-Controlled Topic-Aware Neural Response Generation for Open-Domain Dialog Systems. Inf. Process. Manag., 58, 102392.

[11] Babakov, N., Logacheva, V., Kozlova, O., Semenov, N., & Panchenko, A. (2021). Detecting Inappropriate Messages on Sensitive Topics that Could Harm a Company's Reputation. arXiv preprint arXiv:2103.05345.

[12] Parker, J. (n.d.). How does ChatGPT handle offensive or inappropriate content? James Parker. R etrieved[2024.9.12],from https://www.jamesparker.dev/how-does-chatgpt-handle-offensive-or- inappropriate-content

[13] Chang, J.P., & Danescu-Niculescu-Mizil, C. (2019). Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop. *Conference on Empirical Methods in Natural Language Processing*.

[14] PRENDINGER H, ISHIZUKA M. The empathic companion:a character- based interface that addresses users affective states[J]. Applied Artificial Intelligence, 2005, 19(3/4): 267-285.