A Comprehensive Investigation of Leveraging Bandit Algorithms in Advancing Dialogue Systems

Zesong Guo

The Department of Computer Science, University of Wisconsin-Madison, Madison, 53706, United States of America
Corresponding author's e-mail address: zguo283@wisc.edu

Abstract. Dialogue systems have become integral to modern human-computer interactions, powering applications ranging from virtual assistants to customer support. A critical aspect of these systems is the selection of appropriate responses, a task that is challenging due to the need for real-time adaptation and personalization. While conventional approaches have employed supervised learning, their limitations in handling dynamic interactions and individual preferences have driven the exploration of alternative methods. This review paper addresses a notable gap in existing literature by comprehensively analyzing the application of Multi-Armed Bandit (MAB) algorithms in dialogue response selection. The primary objective is to showcase the potential of MAB algorithms in enhancing dialogue systems' response quality and adaptability. This review presents an in-depth examination of how MAB algorithms operate as part of dialogue systems, focusing on their ability to balance exploration and exploitation to improve response selection. This review surveys recent advancements in dialogue response selection, with specific emphasis on MAB algorithms' integration into this process. Comparative analysis with traditional supervised learning approaches reveals the strengths and limitations of MAB algorithms, highlighting their potential to revolutionize dialogue systems. The findings underscore the capability of MAB algorithms to create context-aware, personalized, and efficient dialogue systems. This review concludes by emphasizing the pivotal role of MAB algorithms in shaping the future of conversational agents, offering a pathway toward more natural and engaging interactions. This review serves as a comprehensive resource for researchers and practitioners seeking to harness the power of MAB algorithms in advancing dialogue systems' capabilities.

Keywords: Multi-armed bandit algorithm, ETC, UCB, TS, Dialogue systems

1. Introduction

Dialogue systems have become an essential part of people's daily interactions with technology, offering human-like conversations and facilitating various applications, such as customer support, virtual assistants, and recommendation systems [1]. Central to the design of these systems is the challenge of selecting appropriate responses that effectively address user inputs while maintaining naturalness and relevance. Traditional approaches have relied on offline supervised learning with extensive context-response datasets. However, such methods struggle to handle real-time interactions and lack personalization, hindering the potential of dialogue systems.

There has been a significant amount of previous research on dialogue systems such as Rule-based systems, Statistical models and Reinforcement learning. While existing literature has extensively explored dialogue systems and response selection, there is a notable gap in the comprehensive review of Multi-Armed Bandit (MAB) algorithms' applications in this domain. This paper seeks to bridge this gap by presenting a detailed analysis of how MAB algorithms have been utilized to optimize dialogue response selection. By understanding the strengths and limitations of these

approaches, researchers and developers can make informed decisions when designing and implementing dialogue systems.

This review paper aims to explore and analyze the novel application of MAB algorithms in dialogue response selection. Unlike traditional supervised learning, MAB algorithms offer an online learning framework that allows dialogue systems to adapt in real-time based on user interactions. The key principle behind MAB algorithms is the exploration-exploitation trade-off, enabling the system to simultaneously explore different response options and exploit promising ones. This adaptive approach holds the promise of enhancing the naturalness and usability of conversational agents.

The primary objective of this review is to provide a comprehensive analysis of MAB-based dialogue systems, focusing on their role in response selection. It aims to examine how MAB algorithms address the challenges of dialogue response selection, such as handling noisy feedback, balancing exploration and exploitation, and adapting to dynamic conversational contexts. By gaining insights into the effectiveness of MAB-based approaches, the development of more sophisticated and proactive dialogue systems can be advanced.

The motivation behind this review lies in the potential impact of MAB algorithms on dialogue systems. By harnessing the power of online learning, dialogue systems can offer more personalized, context-aware, and efficient responses, leading to improved user experiences. Additionally, as dialogue systems become increasingly prevalent in diverse domains, it becomes crucial to explore novel approaches that can address the limitations of traditional supervised learning methods.

In the following sections, this review will first provide an overview of MAB algorithms, their significance to dialogue systems, and the challenges associated with response selection. The theoretical foundations of multi-armed bandit algorithms will be further examined, explaining their exploration-exploitation trade-off and their relevance to dialogue systems. Subsequently, a comprehensive review of methods other than MAB algorithms such as deep will be provided, highlighting their contributions and findings.

2. Applications and methods

2.1. Introduction of MAB

MAB algorithms, or multi-armed bandit algorithms, address the challenge of decision-making in situations where an agent must select from a set of actions with uncertain outcomes to maximize cumulative rewards. This trade-off between exploration and exploitation is a fundamental concept in MAB problems. The application of MAB algorithms finds resonance in dialogue systems, computer programs designed for human interaction through natural language. Dialogue systems encompass open-domain and task-oriented categories. While open-domain systems engage users in broad conversations, task-oriented systems aid in accomplishing specific objectives, like reservations or orders.

In the context of dialogue systems, the selection of optimal responses from a pool of candidates, based on contextual cues and user intent, mirrors a MAB dilemma. Here, candidate responses represent actions, and user satisfaction serves as the reward. MAB algorithms equip dialogue systems to refine responses through interactions, continuously enhancing response quality. For dialogue response selection, a common MAB algorithm is Upper Confidence Bound (UCB), which balances exploration and exploitation by choosing the action with the highest upper bound of the expected

reward. UCB can be applied to select the best response among a set of candidates, based on the user's feedback or some other reward signal. Algorithm 1 provides a simple version of UCB algorithm.

Algorithm 1 UCB

Input: k and δ

for t $\in 1, \ldots,$ n **do**

Choose action $A_t = \text{argmax}_i \text{UCB}(t-1, \delta)$

Observe reward X_t and update upper confidence bounds

end for

Where δ is the error probability and is between 0 and 1.

Furthermore, MAB algorithms extend their utility to multi-domain dialogue systems, where the challenge lies in seamlessly transitioning between diverse domains or tasks. This adaptability necessitates learning domain-switching strategies. MAB algorithms facilitate this by guiding systems to decide whether to persist with a current domain or venture into unexplored ones. For multi-domain dialogue systems, a suitable MAB algorithm is Explore-Then-Commit (ETC), which explores all actions for a fixed number of rounds and then commits to the best action for the remaining rounds. ETC can be used to learn how to switch between different domains and tasks, based on the user's input and the dialogue history. For example, Huang et al. used ETC to select the best domain-specific state generator for dialogue state tracking, based on the domain similarity and accuracy [2]. Algorithm 2 provides the procedure of ETC algorithm.

Algorithm 2 ETC

Input: m

In round t choose action

$$A_t = (t \bmod k) + 1, \quad \text{if } t \leq mk;$$

$$argmax_i \hat{\mu}_i(mk), \quad \text{if } t > mk$$

Where m is the number of times it explores each arm, k is the number of actions, and $\hat{\mu}_i(t)$ is the average reward received from arm I after round t.

A significant stride in dialogue systems involves proactivity, where systems initiate conversations and lead interactions. Proactive dialogue systems can offer valuable insights, recommendations, or reminders, enhancing user experiences. The integration of MAB algorithms empowers systems to discern optimal moments for proactivity, balancing user preferences with non-intrusive engagement. For proactive dialogue systems, a possible MAB algorithm is Thompson Sampling, which is a Bayesian approach that samples an action from the posterior distribution of its reward. Thompson Sampling can be used to learn when and how to initiate or lead the conversation, based on the user's profile, behavior, and feedback. For example, Zhang et al. used Thompson Sampling to select the best proactive dialogue strategy for conversational recommendation, based on the user's preferences and ratings [3]. Algorithm 3 provides the procedure of Thompson sampling algorithm.

Algorithm 3 Thompson sampling

Input: Prior cumulative distribution functions $F_1(1), \ldots, F_k(1)$ for the mean rewards of arms 1,…,k.

In round t choose action

A_t = (t mod k) + 1, if t ≤ mk;

$argmax_i \hat{\mu}_i(mk)$, if t > mk

for $t \in 1, \ldots, n$ **do**

Sample $\theta_i(t) \sim F_i(t)$ independently for each arm i

Choose $\boldsymbol{A_t = argmax_i \theta_i(t)}$

Observe $\boldsymbol{X_t}$ and update the distribution of the arm selected in step 4, i.e.,

$$F_{A_t}(t+1) = UPDATE\left(F_{A_t}(t)\right)$$

end

Where $\boldsymbol{X_t}$ is the reward received at round t.

2.2. Dialogue response selection

In the pursuit of enhancing dialogue response selection, recent studies have introduced innovative methodologies that augment the capabilities of matching models. A notable example is the work by Liu et al. [4]. This study proposes a learning framework that operates on a dynamic "easy-to-difficult" curriculum, encompassing both corpus-level (CC) and instance-level curricula (IC). The CC progressively equips the model with the acumen to discern matching cues within dialogue contexts and response candidates. In parallel, the IC sharpens the model's ability to identify conflicting information between the context and candidate responses. This dual-curriculum approach demonstrates significant performance gains across a spectrum of evaluation metrics.

Another compelling study [5], addresses the challenge of selecting optimal responses from vast corpora or nonparallel datasets. This research pioneers a dense retrieval model supplemented by an interaction layer, bolstered by meticulously designed learning strategies. Impressively, the dense retrieval model outperforms pipeline-based methodologies that combine recall and expressive re-rank modules. Furthermore, the inclusion of nonparallel corpora in the candidate pool enriches response quality, showcasing the potential of dense retrieval models in transforming dialogue systems.

2.3. Pro-activity dialogue systems

Proactivity in dialogue systems refers to the ability of a system to initiate and guide a conversation towards a specific goal or outcome, rather than simply responding to user input. Proactive dialogue systems can be regarded as a form of artificial intelligence that can anticipate user needs, preferences, and intentions based on context, history, and other relevant factors. The fundamental underpinning of proactive dialogue systems revolves around the optimization of user engagement by mitigating cognitive strain, uncertainty, and discontent that commonly accompanies traditional reactive dialogue systems.

Proactive dialogue systems work by using various techniques such as natural language understanding, machine learning, decision making, and planning to generate proactive responses

that are relevant, timely, and informative [6]. For example, a proactive dialogue system can employ context information such as time, location, weather, and user profile to suggest relevant topics or actions before the user asks for. Feedback mechanisms contribute to system refinement, enabling adaptation and optimization over time.

A survey on proactive dialogue systems revealed some of the difficulties in creating such systems, including: 1) Proactivity requires the system to foresee user needs, preferences, and intentions based on context, history, and other pertinent factors. This calls for the design of effective proactivity strategies. However, finding the right balance between being overly proactive (i.e., intrusive) and overly reactive (i.e., passive) can be challenging [6]. 2) Proactivity necessitates that the system be able to handle ambiguity and uncertainty in user input, context, and goals. For instance, based on the user's speech or behaviour, the system might need to infer information about their mood, personality, or culture [6]. 3) Privacy and security protection: Being proactive may entail gathering and handling private data about the user, such as their location, health status, or financial information. As a result, it's critical to confirm that the system complies with moral and legal requirements for data security and privacy [6]. 4) Adapting to dynamic environments: Being proactive may call for the system to change how it behaves in response to new users, tasks, or domains. Therefore, it is crucial to create methods that are scalable and adaptable for learning from data and making generalizations [6].

2.4. Multi-domain Dialogue Systems

The emergence of multi-domain dialogue systems heralds a new era in conversational AI, facilitating seamless transitions between diverse topics within a single conversation [2]. These systems empower users to navigate between domains without compromising coherence, enriching the conversational experience manifold [7]. A natural extension of task-oriented dialogue systems, multi-domain systems excel in handling varied user intents through a combination of natural language understanding, dialogue management, knowledge representation, and machine learning [8].

The study of multi-domain dialogue systems involves various research topics such as natural language understanding, dialogue management, knowledge representation, and machine learning [9]. The main module of multi-domain dialogue systems is the dialogue manager, which is responsible for selecting the appropriate response based on the user input and the system state [10]. The workflow of multi-domain dialogue systems typically involves several stages such as intent recognition, slot filling, context tracking, policy learning, and response generation [2].

3. Discussion

3.1 Dialogue response selection

The task of dialogue response selection holds paramount importance in the development of natural and engaging conversational agents. Deep learning models have emerged as a powerful tool for this task, offering a spectrum of advantages. These models excel at capturing intricate semantic and syntactic features from raw text, enabling them to grasp the nuances of user intent and context. Furthermore, their ability to handle large-scale and diverse datasets ensures the versatility necessary for real-world applications. Leveraging pre-trained language models enhances generalization and robustness, allowing dialogue systems to offer more contextually appropriate responses. However, there are inherent challenges within the realm of dialogue response selection using deep learning

models. The acquisition of substantial annotated data remains a significant requirement, potentially impeding the scalability of these approaches. Data sparsity and imbalance can also hamper model performance, leading to biased or suboptimal responses. In contrast, MAB algorithms, epitomized by the Upper Confidence Bound (UCB) approach, strike an equilibrium between exploration and exploitation. UCB dynamically balances the quest for optimal responses, drawing from a calculated upper bound of expected rewards. This equips dialogue systems with the means to explore various response candidates while consistently exploiting the best available option, a synergy that is crucial for enhancing user satisfaction and engagement.

The integration of MAB algorithms into dialogue response selection offers a promising avenue to surmount challenges. The utilization of pre-trained language models and the incorporation of external knowledge, aligned with MAB-driven decision-making, elevate the quality of responses. By fostering diversity and personality in generated replies, dialogue systems can emulate natural human conversations, a vital step toward seamless human-machine interactions.

3.2 Proactive dialogue systems

Proactivity within dialogue systems represents a transformative leap, ushering in a new era of user interaction. MAB algorithms, such as the Bayesian-based Thompson Sampling, provide the essential underpinning for this shift. Thompson Sampling empowers systems to learn and optimize proactive strategies by sampling actions from posterior reward distributions [11]. This ability to strike a balance between initiating conversations and respecting user preferences underpins the success of proactive dialogue systems. The integration of MAB algorithms imbues proactive dialogue systems with unprecedented capabilities. The nuanced understanding of user behavior, combined with proactive initiation, results in more meaningful and contextually relevant interactions.

Despite its potential, proactive dialogue systems encounter challenges rooted in the complexities of reward design and the exploration-exploitation trade-off. Ethical and social considerations loom large, demanding careful navigation to avoid manipulation or deception of users. The transparency and explainability of such systems are also pivotal aspects warranting future exploration. In the future, the integration of user feedback or reinforcement signals could refine proactive dialogue strategies, aligning system actions more closely with user preferences. Fostering user trust and engagement remains a priority, with potential implications for system design and interaction mechanisms. Rigorous evaluation methodologies encompassing both system performance and user experience will shape the evolution of proactive dialogue systems.

3.3 Multi-domain dialogue systems

In the realm of multi-domain dialogue systems, MAB algorithms offer a crucial bridge to effective domain handling and seamless transitions. The Explore-Then-Commit (ETC) strategy, a representative MAB paradigm, lends itself to the dynamic world of multi-domain interactions. ETC, characterized by initial exploration followed by focused exploitation, offers a strategic framework for dialogue systems to deftly switch between domains based on user inputs and context.

MAB-driven multi-domain dialogue systems hold the potential to revolutionize user experiences. The incorporation of shared and domain-specific representations, fortified by MAB-based adaptation mechanisms, ensures coherent and relevant responses. As the complexity of integrated domains grows, MAB algorithms, operating in tandem with domain adaptation techniques, promise to mitigate challenges of scalability and domain interference. The quest for

consistent and user-centric multi-domain interactions aligns with MAB's core principles, driving the evolution of these systems toward heightened coherence and efficiency.

In this dynamic landscape, the application of MAB algorithms emerges as a formidable contender, offering distinct advantages over deep learning approaches. While deep learning models harness their prowess in semantic comprehension, MAB algorithms provide a systematic mechanism for addressing exploration and exploitation trade-offs. Deep learning models often require substantial annotated data, a hurdle that MAB algorithms mitigate through adaptive learning. Nevertheless, MAB algorithms introduce their own set of complexities, particularly in reward design and balancing exploration with exploitation. In this intricate interplay between MAB algorithms and deep learning, the future trajectory of dialogue systems is being charted. As dialogue systems continue to evolve, the harmonious fusion of these methodologies holds the key to crafting more intelligent, adaptable, and user-centric conversational agents.

4. Conclusion

The convergence of MAB algorithms and dialogue systems has ushered in a revolutionary era in artificial intelligence. MAB algorithms, strategically designed to navigate the intricate balance between exploration and exploitation, have introduced a structured approach to tackle the complexities of dialogue response selection, proactive dialogue systems, and multi-domain interactions. Through the lens of MAB algorithms, the art of dialogue response selection unfolds, demanding optimal choices from a pool of potential responses. While deep learning models excel in semantic comprehension, they heavily rely on annotated data. In contrast, MAB algorithms like UCB excel in decision-making amid uncertainty. UCB empowers dialogue systems to explore diverse responses while consistently exploiting the most promising ones, elevating user satisfaction and engagement. The integration of MAB algorithms extends to proactive dialogue systems, where conversational agents initiate and steer interactions. Proactive strategies, anchored by Thompson Sampling, a Bayesian MAB approach, achieve a delicate balance between user engagement and autonomy. However, as proactive systems advance, ethical considerations and transparency gain prominence, with MAB algorithms guiding thoughtful and user-centric interactions. In the realm of multi-domain dialogue systems, MAB algorithms, exemplified by the ETC approach, prove invaluable in shaping domain transitions. MAB-driven adaptation mechanisms facilitate seamless multi-domain interactions, ensuring users traverse diverse topics effortlessly. While MAB algorithms offer unique advantages over deep learning, complexities arise in reward design and trade-off management. The fusion of MAB algorithms and deep learning paves the way for adaptive conversational agents that redefine human-computer interactions. This amalgamation harmonizes the strengths of both methodologies, promising enhanced user experiences in the evolving landscape of artificial intelligence.

5. References

[1] Bouneffoud D and Rish I 2019 A Survery on Practical Application of Multi-Armed and Contextual Bandits (NY USA: IBM Thomas J. Watson Research Center Yorktown Heights)

[2] Huang Y Feng J and Hu M et al 2020 Meta-Reinforced Multi-Domain State Generator for Dialogue Systems (China: Mobile Research, JIUTIAN Team)

[3] Zhang Y M Wu L F and Shen Q et al. 2021 Multi-Choice Questions based Multi-Interest Policy Learning for Conversational Recommendation (China: Tongji University, USA: JD Silicon Valley Research Center)

[4] Su Y X Cai D and Zhou Q Y et al. 2021 Dialogue Response Selection with Hierarchical Curriculum Learning (Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing) p1740 - 1751
[5] Lan T Cai D and Wang Y et al. 2022 Exploring Dense Retrieval for Dialogue Response Selection (China: Beijing Institute of Technology/The Chinese University of Hong Kong UK: University of Cambridge)
[6] Deng Y Lei W Q and Lam W et al. 2023 A Survey on Proactive Dialogue Systems: Problems, Methods, and Prospects (China: The Chinese University of Hong Kong/Sichuan University, Singapore: National University of Singapore)
[7] Kung P N Chang C C and Yang T H et al. 2021 Multi-Task Learning for Situated Multi-Domain End-to-End Dialogue Systems (Taiwan: National Taiwan University)
[8] Tagge C 2006 A Multi-Domain Dialogue Management System (USA: Stanford University)
[9] Nishimoto B E and Costa A H R 2021 Slot Sharing Mechanism in Multi-domain Dialogue Systems (Brazil: University of Sao Paulo)
[10] Wang K Tian J F and Wang R et al. 2020 Multi-Domain Dialogue Acts and Response Co-Generation (China: Sun Yat-sen University/Alibaba Group)
[11] Kim G S and Paik M C 2019 Contextual Multi-Armed Bandit Algorithm for Semiparametric Reward Model (Korea: Seoul National University)