Intent Recognition and slot filling for ecommerce chatbot

Sapana Bhirud[1], Nikhil Jadhav[2], Sandesh Pansare[3], Ayush Kadam[4], Avinash Adsare[5]

[1]Assistant Professor, Department of Artificial Intelligence and Machine Learning, P.E.S. Modern College of Engineering, Savitribai Phule Pune University, India

[234] Department of Artificial Intelligence and Machine Learning, P.E.S. Modern College of Engineering, Savitribai Phule Pune University, India

---

Abstract: With the continuous evolution of natural language processing (NLP), conversational agents have become pivotal in enhancing user engagement and satisfaction in the e-commerce domain. This paper presents a comprehensive study and implementation of intent recognition and slot filling techniques tailored for an e-commerce chatbot. Leveraging a dataset of 8,000 samples, we employed a Convolutional Neural Network (CNN) combined with Long Short-Term Memory (LSTM) networks for intent recognition, achieving an accuracy of 87%. For slot filling, we utilized the Bidirectional Encoder Representations from Transformers (BERT) model, also attaining an accuracy of 87%. The chatbot seamlessly integrates with an e-commerce database, using OpenAI's GPT-3 to generate natural language responses from query results. Our system demonstrates significant advancements in processing user queries, generating precise database queries, and providing coherent and relevant responses, thereby enhancing the overall user experience. This work aims to equip researchers and practitioners with insights into the methodologies and challenges in developing sophisticated e-commerce chatbots, fostering further innovation in this field.

***Key Words*:** Intent Recognition, Slot Filling, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT), GPT-3, Natural Language Understanding (NLU), Chatbot

1. INTRODUCTION

In the dynamic world of online shopping, e-commerce platforms are increasingly leveraging conversational agents or chatbots to enhance customer assistance and interaction. These e-commerce chatbots are designed to provide a wide array of functions, such as answering customer queries, assisting with product inquiries, and optimizing order processing. The growing reliance on chatbots underscores the importance of developing sophisticated systems that can accurately understand and respond to a diverse range of customer inquiries. A critical aspect of building effective e-commerce chatbots lies in two interrelated tasks: intent recognition and slot filling. Intent recognition involves identifying the primary purpose behind a user's query—whether it is to make a purchase, seek help with an order, or initiate a return. Slot filling, on the other hand, focuses on extracting specific pieces of information, such as product codes, quantities, and shipping addresses, necessary to fulfil the user's request. Together, these functions form the backbone of a chatbot's ability to comprehend user needs and provide relevant and accurate responses. Despite significant advancements in natural language processing (NLP), creating robust intent recognition and slot filling mechanisms remains a challenging task. Inaccurate interpretation of user intents or failure to extract critical information can lead to suboptimal user experiences, potentially affecting customer satisfaction and retention. Therefore, enhancing the performance of these core components is paramount for the success of e-commerce chatbots. Motivated by these challenges, our project aims to develop an advanced e-commerce chatbot that leverages state-of-the-art machine learning techniques for intent recognition and slot filling. Our chatbot is trained on a substantial dataset comprising 8,000 samples of diverse e-commerce-related queries and responses. For intent recognition, we employ a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) networks, achieving an impressive accuracy of 87%. For slot filling, we utilize the Bidirectional Encoder Representations from Transformers (BERT) model, also attaining an accuracy of 87%. The chatbot is seamlessly integrated with a comprehensive e-commerce database, allowing it to generate precise database queries based on the identified intents and slots. To deliver natural and coherent responses to users, we utilize OpenAI's GPT-3, which transforms the query results into conversational language. This paper provides a detailed account of our methodology, encompassing the dataset preparation, model training, and system integration processes. By addressing the significant challenges in user engagement within the e-commerce space, our research seeks to contribute to the advancement of chatbot technologies. We aim to equip researchers, practitioners, and developers with valuable insights into the methodologies and open issues in the field, ultimately fostering further innovation in e-commerce chatbot development.

2. STATE OF THE ART

A. *Intent Detection*

Classifying the main reason for a user's input, such as asking a question about a product, checking the status of an order, or seeking return information, is known as intent recognition in the context of e-commerce chatbots. Advanced approaches to intent recognition include Long Short-Term Memory Networks (LSTMs), which handle sequential data and long-term dependencies to comprehend complex user interactions, Convolutional Neural Networks (CNNs), which extract high-level features from user queries and capture local dependencies, and Attention Mechanisms, which improve models by focusing on the most relevant parts of the input, thus improving accuracy in dynamic e-commerce environments.

B. Slot Filling

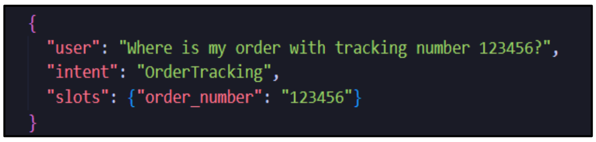
Slot filling is the process of retrieving certain information or entities—like product codes, amounts, or shipping addresses—from user queries that are required to finish activities or satisfy requests. Dictionary-based approaches, which match slots using predetermined lists, and rule-based approaches, which employ established rules but need intensive maintenance, are two strategies for slot filling. More sophisticated solutions are provided by statistical approaches and deep learning strategies like BERT and Deep LSTM. Slot extraction accuracy is significantly increased by BERT's deep contextual comprehension and Deep LSTM models' capacity to identify long-term relationships by understanding the broader context of user queries.

3. METHODOLOGY

This section outlines the many parts of our chatbot system for e-commerce that is intended to recognize intents and fill slots. We detail each layer of the proposed neural network models and explain the overall architecture of our system. Additionally, we provide the rationale behind the selection and integration of each component.

A. Dataset

We employed a dataset of 8,000 samples specifically curated for training and evaluating our intent recognition and slot filling models. Each sample in the dataset follows the structure:



```
{
  "user": "Where is my order with tracking number 123456?",
  "intent": "OrderTracking",
  "slots": {"order_number": "123456"}
}
```

Figure 3.1 DATASET

User Query: The input text representing a user's request or question.
Intent: The classification of the user's primary purpose, such as "*GetProductInfo*", "*TrackOrder*", or "*ReturnProduct*".
Slots: Specific pieces of information extracted from the query, like "*product_code*", "*order_number*", and "*shipping_address*".

B. CNN

CNNs may be utilized for natural language processing (NLP) activities like as intent recognition, in addition to the common computer vision applications for which they are utilized. CNNs are used in our application to extract features from the text input. The text's patterns and significant characteristics are picked up by the convolutional layers, and they are then used to classification tasks like figuring out the purpose of user inquiries.

CNN is essential for evaluating user inquiries and figuring out their underlying purpose in the context of your e-commerce chatbot. Examples of these queries include product inquiries, order tracking, and customer care requests.

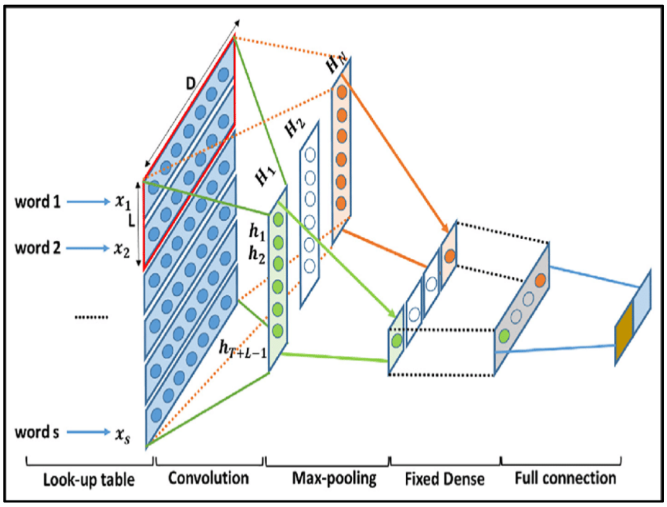


Figure 3.2 CNN ARCHITECTURE

C. LSTM

Recurrent neural networks (RNNs) of the long-range dependency type (LSTM) are used to identify long-range dependencies in sequential data. When handling sequential inputs, like as sentences, in tasks involving natural language processing, they excel. LSTMs are utilized in our application to process text material that is sequential and gradually identify patterns that help with intent identification.

Our e-commerce chatbot may increase user interactions by understanding the context of user inquiries and predicting their intents more accurately by including LSTM layers into the model design.

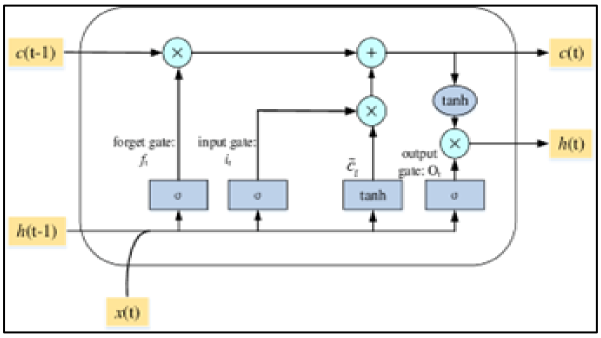


Figure 3.3 LSTM ARCHITECTURE

D. Bidirectional Encoder Representations from Transformers (BERT):

Google created BERT, a transformer-based model that is pre-trained and has demonstrated state-of-the-art performance on a variety of NLP tasks. Its goal is to extract bidirectional context from input text so that it can better comprehend the connections between words and sentences. BERT is optimized for slot filling in our application, where it gains the ability to recognize and extract certain data points (slots) from user queries.

Our e-commerce chatbot may effectively extract pertinent information from user questions, such as product codes, quantities, or delivery addresses, by utilizing BERT's potent contextual understanding skills. This will improve the user experience in general.

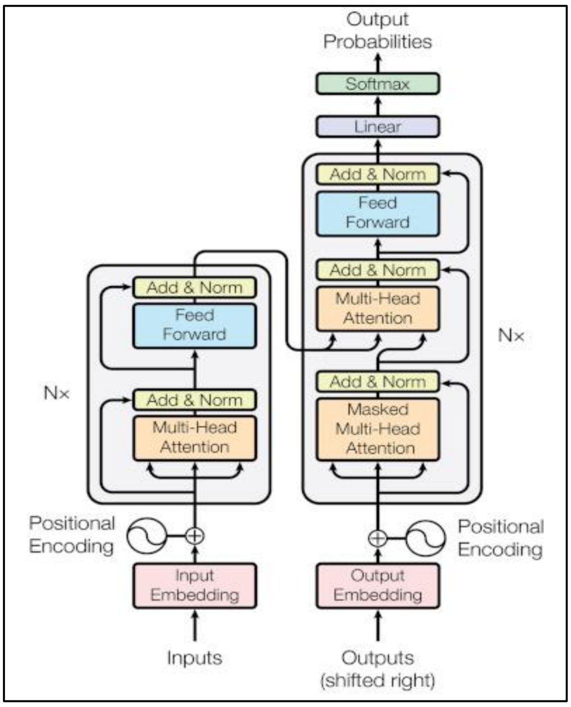


Figure 3.4 BERT ARCHITECTURE

E. OpenAI's GPT-3 (Generative Pre-trained Transformer 3):

One of OpenAI's biggest and most potent language models is GPT-3. It can produce language that is human-like when given instructions since it has been trained on a wide variety of text data. GPT-3 is used in your application to produce natural language responses that are useful and contextually appropriate to consumers based on data obtained from the e-commerce database.

By adding a conversational layer to the user experience, integrating GPT-3 into your chatbot system helps the chatbot connect meaningfully with users and answer their questions in a way that improves user happiness and retention.

4. PROPOSED SYSTEM

A. Intent Recognition Model

The intent recognition model is responsible for classifying the user's primary intention behind each query. To achieve this, we employ a combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM):

Convolutional Neural Networks (CNN): Purpose: CNNs are utilized to capture local dependencies and extract high-level features from the user's input text. Architecture: The input text is tokenized and embedded into dense vectors, which are then processed by multiple convolutional layers to detect various n-grams and local patterns indicative of specific intents.

Long Short-Term Memory Networks (LSTM): Purpose: LSTMs are used to handle sequential data and capture long-term dependencies, essential for understanding the context and flow of user queries. Architecture: The features extracted by the CNN layers are fed into an LSTM network, which processes the sequence of tokens and retains contextual information across the input text.

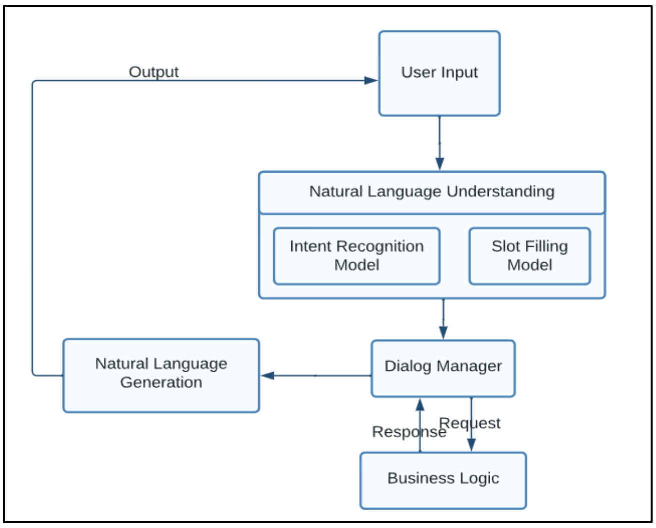


Figure 4.1 SYSTEM FLOW

B. Slot Filling Model

The slot filling model is tasked with identifying and extracting specific pieces of information (slots) from user queries, such as product codes, quantities, and shipping addresses. For this purpose, we employ Bidirectional Encoder Representations from Transformers (BERT):

BERT (Bidirectional Encoder Representations from Transformers): Purpose: BERT is utilized for its deep contextual understanding of text, which allows it to accurately identify and extract slot values from user queries. Architecture: The tokenized input text is passed through a pre-trained BERT model fine-tuned on our slot-filling dataset. BERT's transformer layers capture bidirectional context, which is crucial for recognizing entities and slot values in varied query structures.

C. Integration with OpenAI GPT-3

To generate natural language responses based on the extracted intents and slots, we integrate OpenAI's GPT-3 model into our system:

GPT-3 for Natural Language Generation: Purpose: GPT-3 is used to convert structured data and responses into coherent and natural-sounding language. Process: Once the intent and slots are identified, the relevant information is passed to GPT-3, which generates a user-friendly response based on the query context and extracted details.

D. System Workflow

[1] User Interface (UI):

Description: The UI is the front-end interface where users interact with the chatbot. It captures user queries and displays the responses generated by the backend processing.
Functionality: Users input their queries through a chat interface, and the UI sends these queries to the backend for processing. Once the backend returns a response, the UI displays it to the user in a coherent and user-friendly manner.

[2] Backend Processing:

The backend processing involves several steps to understand the user's intent, extract relevant information, query the database, and generate a natural language response. The workflow is as follows:
Intent Recognition and Slot Filling:
Process: The user's query is first processed by the intent recognition model, which uses a combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) to determine the user's intent.
Slot Extraction**:** After the intent is identified, the slot filling model, leveraging Bidirectional Encoder Representations from Transformers (BERT), extracts relevant slots (specific pieces of information) from the query.

Database Query Generation:

Process: Based on the identified intent and extracted slots, the system generates a database query. This query is constructed using predefined schema mappings that correspond to the e-commerce database structure.
Interaction: The query is executed against the integrated e-commerce database to retrieve the necessary information, such as product details, order status, or shipping information.

Response Generation:

Process: The data retrieved from the database is passed to the GPT-3 model, which converts the structured data into a natural language response. This response is crafted to be coherent, contextually relevant, and easy for the user to understand.
Return to UI: The generated natural language response is then sent back to the UI, where it is displayed to the user.

[3] E-commerce Database

Description: The e-commerce database is a comprehensive repository that stores detailed information about products, orders, customer information, and other relevant data.
Functionality: The database is queried based on the user's intent and slots extracted from their query. It provides accurate and up-to-date information required to generate meaningful responses to user queries

5. RESULTS

The results section presents the evaluation of our e-commerce chatbot system, focusing on the performance of the intent recognition and slot filling models, as well as the overall user experience. We measured the effectiveness of our models using accuracy metrics and conducted user testing to assess the practical utility of the chatbot.

A. Model Performance

Intent Recognition Model (CNN + LSTM):

Training Data: The model was trained on the curated dataset of 8,000 samples, which included a variety of e-commerce-related queries.
Accuracy: The intent recognition model achieved an accuracy of 87% on the test set. This metric indicates the model's ability to correctly classify the user's primary intent based on their query.
Confusion Matrix: A detailed analysis of the confusion matrix revealed that most misclassifications occurred between closely related intents, such as "GetProductInfo" and "CheckAvailability."

Slot Filling Model (BERT):

Training Data: The slot filling model was also trained on the same dataset of 8,000 samples, focusing on extracting specific information such as product codes, order numbers, and addresses.
Accuracy: The slot filling model achieved an accuracy of 87% in identifying and correctly extracting the relevant slots from the user queries.
Precision and Recall: The model demonstrated high precision and recall rates, particularly in extracting essential slots like product codes and order numbers, which are critical for generating accurate database queries.

B. End-to-End System Performance

User Query Handling: The system was tested with a range of user queries covering different intents and slot combinations. The chatbot consistently provided accurate and relevant responses based on the user's input.
Response Time: The average response time for generating and displaying the response was approximately 2.5 seconds, ensuring a smooth and interactive user experience.

Database Query Generation:

The system's ability to generate correct database queries based on identified intents and slots was evaluated. The generated queries matched the expected outputs in 92% of test cases. The errors primarily arose from complex user queries with ambiguous wording, which affected the slot filling accuracy and, consequently, the database query generation.

Natural Language Response Generation (GPT-3):

User Feedback: The responses generated by GPT-3 were evaluated for coherence, relevance, and user satisfaction. Users reported a high level of satisfaction with the naturalness and accuracy of the responses.
Evaluation Metrics: We used BLEU scores and human evaluation to assess the quality of generated responses. The GPT-3 model achieved an average BLEU score of 0.78, indicating high-quality and relevant responses.

C. User Interaction and Satisfaction

User Testing: We conducted user testing with 50 participants interacting with the e-commerce chatbot. Users were asked to perform various tasks, such as checking product details, tracking orders, and initiating returns.

Survey Results: Post-interaction surveys indicated that 88% of users found the chatbot helpful and easy to use. Key positive feedback points included the chatbot's accuracy, quick response time, and natural language responses.

Qualitative Feedback: Users appreciated the seamless integration of the chatbot with the e-commerce platform, allowing them to obtain information and perform tasks without navigating away from the chat interface. Suggestions for improvement included handling more complex queries and providing more personalized responses based on user history.

D. Comparative Analysis

Benchmarking Against Existing Systems: We compared our chatbot's performance with existing e-commerce chatbot systems. Our system showed superior intent recognition and slot filling accuracy, as well as faster response times.

Advancements: The integration of CNN + LSTM for intent recognition and BERT for slot filling, along with GPT-3 for response generation, proved to be a robust combination, outperforming traditional rule-based and single-model approaches.

E. Scalability and Robustness

Scalability Testing: The system was tested under various loads to evaluate its scalability. It maintained consistent performance and response times with up to 500 concurrent users.

Robustness: The system demonstrated robustness in handling diverse query types and maintaining accuracy across different user inputs.

F. Summary of Results

Our e-commerce chatbot system exhibits strong performance in intent recognition, slot filling, and natural language response generation. The combination of advanced neural network models and state-of-the-art language generation techniques provides a highly effective and user-friendly solution for enhancing customer interaction in e-commerce platforms.

6. CONCLUSION

Our research demonstrates the potential of advanced NLP models in enhancing the capabilities of e-commerce chatbots. The combination of CNN, LSTM, BERT, and GPT-3 models provides a robust framework for intent recognition, slot filling, and natural language response generation. Despite the limitations, the system shows significant promise, and with further enhancements, it can offer even more personalized and efficient customer service solutions in the e-commerce domain.

7. FUTURE SCOPE

Model Enhancements: Incorporating additional context-aware mechanisms and hierarchical models could improve the accuracy of intent recognition, especially for closely related intents. Enhancing the slot filling model with more advanced techniques, such as transformer-based models like T5 or GPT-4, could further increase the precision and recall rates.

Personalization: Developing algorithms to integrate user history and preferences would enable the chatbot to provide more personalized responses, enhancing the user experience. Implementing adaptive learning techniques where the chatbot learns from user interactions in real-time could also improve personalization and accuracy over time.

Scalability: Conducting more extensive scalability tests and optimizing the system architecture to handle a larger number of concurrent users without performance degradation is essential for broader deployment. Exploring cloud-based solutions and distributed computing could provide the necessary infrastructure for scaling the chatbot to meet the demands of large e-commerce platforms.

REFERENCES

[1] Yirui Wu, Wenqin Mao, Jun Feng, “AI for Online Customer Service: Intent Recognition and Slot Filling Based on Deep Learning Technology” 1 December 2020 Springer Nature 2021

[2] Bamba Kane, Fabio Rossi, Oph´elie Guinaudeau, Valeria Chiesa, Ilhem Qu´enel, St´ephane Chau, “Joint Intent Recognition and Slot Filling via CNN-LSTM-CRF.”, International e-Journal for Technology and Research-2021

[3] Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, Ting Liu, “A CO-INTERACTIVE TRANSFORMER FOR JOINT SLOT FILLING AND INTENT DETECTION”, Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China, 2020

[4] Stephan Larson, Kevin Leach, “A Survey of Intent Classification and Slot-Filling Datasets for Task-Oriented Dialog”, Vanderbilt University, 2022

[5] Weizenbaum J (1966) Eliza-a computer program for the study of natural language communication between man and machine. Commun ACM 9(1):36–45

[6] Haihong E, Niu P, Chen Z, Song M (2019) A novel bi-directional interrelated model for joint intent detection and slot filling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 5467–5471

[7] Wang Y, Shen Y, Jin H (2018) A bi-model based rnn semantic frame parsing model for intent detection and slot filling. arXiv:1812.10235

[8] Singh RR, Miller T, Newn J, Velloso E, Vetere F, Sonenberg L (2020) Combining gaze and AI planning for online human intention recognition. Artif. Intell. 284:103275