Lewis Hamilton Chatbot Report and Evaluation

Sree Harsha Koyi

SXK230025

This system describes a comprehensive chatbot framework designed to process and respond to user inputs in an intelligent and context-aware manner. The system integrates various Natural Language Processing (NLP) techniques and libraries, including NLTK (Natural Language Toolkit), Spacy, and Scikit-learn, to preprocess text, understand user queries, and retrieve or generate relevant responses. Below is a detailed report of the system's components, NLP techniques utilized, and their specific applications within the framework.

System Overview

The chatbot system is structured to offer dynamic interactions with users, capable of understanding and storing user preferences, performing web lookups for queries beyond its knowledge base, and employing machine learning to enhance response relevance over time. The system architecture is designed to be modular, with components for preprocessing text, managing a knowledge base, performing web lookups, and handling user models.

NLP Techniques and Libraries

1. Text Preprocessing

Text preprocessing is a critical step in the system, involving tokenization, stop word removal, and lemmatization to clean and normalize user input and knowledge base content for further processing. This step employs NLTK for tokenization and stop word removal, alongside the WordNetLemmatizer for lemmatization, preparing text for feature extraction and similarity analysis.

2. Feature Extraction

The TfidfVectorizer from Scikit-learn is used to convert the preprocessed text into a TF-IDF (Term Frequency-Inverse Document Frequency) matrix. This technique helps in understanding the importance of words within documents in the knowledge base relative to their frequency across all documents, enabling effective similarity comparisons.

3. Cosine Similarity

To identify the most relevant response to a user query, the system employs cosine similarity measures between the TF-IDF vector of the user's input and the vectors of sentences in the knowledge base. This approach allows the chatbot to find the knowledge base entry that best matches the query based on textual similarity.

4. Named Entity Recognition (NER)

Spacy's NLP library is utilized for Named Entity Recognition to extract personal information (such as names, organizations, locations) from user inputs. This information is used to personalize interactions and update the user model, enhancing the chatbot's ability to provide context-aware responses.

5. User Model Management

The system manages personalized user models stored in JSON format, containing information such as personal details, likes, dislikes, and interaction history. These models enable the chatbot to tailor responses based on individual user preferences and history.

6. Web Lookup

For queries that cannot be answered using the knowledge base, the system performs web lookups using the Requests library to fetch information from Google search results. BeautifulSoup is then used to parse and extract relevant snippets from the HTML content.

System Workflow

Initialization: On startup, the system checks for the existence of a knowledge base and user models directory, creating them if they do not exist.

User Interaction: The chatbot prompts for user input, which undergoes preprocessing, entity extraction, and analysis against the knowledge base or web, as appropriate.

Response Generation: Depending on the query's nature and the available data, the chatbot generates a response either from the knowledge base, through web lookup, or using hardcoded responses for simple greetings or farewells.

User Model Update: After each interaction, the system updates the user model with new information extracted from the input and the query-response pair, enabling continuous learning and personalization.

**Dialog Logic**

The chatbot's dialog logic is primarily based on a combination of hardcoded responses for common greetings or farewells, similarity-based retrieval from a knowledge base for known topics, and dynamic web lookups for queries outside its knowledge base. The decision to use hardcoded responses, knowledge base lookup, or web lookup is determined by the similarity score between the user's query and the knowledge base entries.

This is how decisions are made for each input:

Start

└── Receive User Input

├── Check for Hardcoded Responses

│ ├── If "greeting" or "farewell": Return hardcoded response

│ └── Else: Proceed to Text Preprocessing

│ └── Preprocess Text

│ ├── Tokenize

│ ├── Remove Stop Words

│ └── Lemmatize

│ └── Vectorize Text (TF-IDF)

│ └── Calculate Cosine Similarity with Knowledge Base

│ ├── If similarity high(>0.8)

│ │ └── Return Knowledge Base Entry

│ ├── If similarity low

│ │ ├── Perform Web Lookup

│ │ │ └── Return Web Lookup Result

│ │ └── If user likes known

│ │ └── Generate Personalized Response

│ └── Update User Model (Optional)

│ ├── Extract Personal Information (NER)

│ └── Save Interaction

└── End

└── Present Generated Response to User

**Sample Dialog Interactions**

**Basic Interaction:**

Please enter your user ID to start: 100

Hello! I'm your chatbot. Ask me anything or type 'quit' to exit.

You: hi

Chatbot: Hello there! How can I assist you today?

**Interactions based on Knowledge base:**

You: who is lewis hamilton

Chatbot: Sir Lewis Carl Davidson Hamilton MBE HonFREng (born 7 January 1985) is a British racing driver competing in Formula One, driving for Mercedes, and has also driven for McLaren. Hamilton

You: where does lewis hamilton born

Chatbot: Born and raised in Stevenage, Hertfordshire, he began karting in 1993 at the age of eight and achieved success in local, national and international ...Nicolas Hamilton · List of Formula

**Interactions based on Weblookup:**

You: who is hamilton rival

Chatbot: Max Verstappen

You: how many race wins does lewis hamilton have

Chatbot: His records to date included 103 victories, 103 pole positions, 183 podium appearances and numerous other distinctions .

Screenshots of the responses:

A black screen with many small colored lines

Description automatically generated

A black background with many small colored text

Description automatically generated with medium confidence

**Appendix A: Knowledge Base and Live Lookup**

The knowledge base is a dictionary structure, initialized empty but designed to be populated with serialized objects representing processed information on various topics. A sample entry might look like this: knowledge\_base = {

"natural language processing": ["NLP enables computers to understand human language."]

}

The live lookup function perform\_web\_lookup dynamically queries Google for information not present in the knowledge base, ensuring the chatbot can respond to a wide range of queries.

**Appendix B: Sample User Models**

User models are JSON files storing information about the user, including personal preferences, likes, dislikes, and interaction history. A sample user model might look like:

{

"name": "Sree Harsha",

"personal\_info": {

"PERSON": ["Sree Harsha"],

"ORG": ["UTD"],

"GPE": ["Texas"],

"LOC": ["Dallas"]

},

"likes": ["technology", "science"],

"dislikes": ["fast food"],

"interactions": [

{

"query": "What is AI?",

"response": "Artificial Intelligence is a branch of computer science..."

}

],

"feedback": []

}

Evaluations:

Strengths

**Flexibility:** Can handle a variety of queries through both its knowledge base and web lookup.

**Personalization**: Extracts and utilizes personal information to tailor interactions.

**Learning Capability**: Grows its knowledge base over time based on user interactions.

**Weaknesses**

**Dependency on External Resources:** The web lookup's effectiveness depends on the availability and accessibility of external websites.

**Limited Understanding**: May not fully grasp complex queries or provide depth in answers due to the limitations of TF-IDF and cosine similarity.

**Privacy Concerns**: Collects personal information, necessitating robust data protection measures.

Results of the Survey:

For 3 Participants:A green circle with yellow lines

Description automatically generated

A green circle with a yellow triangle

Description automatically generated

Conclusion

This chatbot system showcases the application of multiple NLP techniques to achieve a responsive, intelligent, and personalized user experience. Through text preprocessing, feature extraction, similarity analysis, NER, and user model management, the system can understand and respond to a wide range of queries while adapting to individual user preferences and expanding its knowledge base over time. While it showcases the potential for personalized and informative dialogue, its reliance on external data and the inherent limitations of its NLP methodologies suggest areas for future improvement, particularly in understanding complex queries and ensuring user privacy.