PharmaSUG China 2025 - Paper AI-128

An Automated Chatbot for Addressing Submission-Related Questions

Linjie Jia, Sanofi Aventis

ABSTRACT

As a statistical programmer in the CoE-esub team, which specializes in submission-related work, I frequently encounter scattered and numerous questions about the submission process. For those unfamiliar with preparing submission packages, knowing where to begin or find relevant information can be challenging. To address this, we developed an intelligent chatbot designed to answer submission-related queries in real time. Leveraging machine learning techniques, the chatbot interprets user questions and retrieves accurate responses from a curated database. This database is built from frequently asked questions and daily work records, with continuous updates based on newly logged issues. This paper discusses the design, implementation, and performance of this chatbot, demonstrating its potential to enhance efficiency in submission package preparation.

INTRODUCTION

Regulatory submission plays a pivotal role in the pharmaceutical industry, requiring adherence to complex standards such as the Clinical Data Interchange Standards Consortium (CDISC) and region-specific regulations, including the Food and Drug Administration (FDA) for US, Pharmaceuticals and Medical Devices Agency (PMDA) for Japan, European Medicines Agency (EMA) for Europe, and National Medical Products Administration (NMPA) for China. While professionals may differ in their direct experience with submission packages, understanding these requirements is essential.

However, preparing submission materials often involves navigating diverse challenges across studies. Traditional approaches, such as manual searches in regulatory databases, are inefficient and prone to errors. This paper introduces a brand-new method involving the development of a Python-based chatbot system designed to respond to submission-related inquiries. The proposed solution aims to streamline administrative processes by providing efficient and accurate automated responses, thereby significantly reducing response times and improving overall operational efficiency.

WHY CHOOSE PYTHON

Python is a high-level, object-oriented programming language that supports structured, functional, and object-oriented programming paradigms. It is also widely used as a scripting language. Compared to SAS and R, Python may not have the same established reputation in specialized statistical analysis domains. However, it offers distinct advantages.

- Python excels in input/output (I/O) operations and provides robust support for multithreading and process management, enhancing its utility for various applications.
- Python's extensive standard library, particularly its rich string handling capabilities and excellent support for regular expressions, makes it exceptionally well-suited for text processing tasks.
- Natural Language Processing (NLP) Libraries: Python has excellent NLP libraries like NLTK, spaCy, Hugging Face Transformers, and TensorFlow/ PyTorch for machine learning.
- Web Frameworks: Flask, FastAPI, and Django make it easy to deploy chatbot backends.
- AI/ML Ecosystem: Comprehensive support for modern AI techniques used in chatbots.

METHODOLOGY

DATA SOURCE AND PREPARATION

The knowledge base for the question-answering system was constructed using two primary data sources:

EXCEL database structure

The core component of the knowledge base consists of a structured Excel database containing question-answer pairs relevant to pharmaceutical submissions. The database is organized into categories including regulatory requirements, document specifications, and some other frequently asked questions (FAQ) in the submission package preparation corresponding work . Each entry in the database contains fields for the question text, answer text, category classification, relevant keywords, and metadata tags for cross-referencing with regulatory documents.

The database was compiled through systematic extraction of common questions and their answers from historical inquiries, regulatory guidance documents, and expert knowledge, covering the most frequently encountered submission-related topics. Figure 1 illustrates a representative structure of the database implemented in Microsoft Excel.

Number	Question	Answers	Category	Key Words	Metadata Tags

Figure 1. The structure of the database in Excel

Regulatory documents Integration

To supplement the structured database, several regulatory PDF documents were processed to extract additional context and metadata. These documents included regulatory guidelines and submission requirements from major regulatory authorities. Text extraction from PDFs was performed using the PyPDF2 library, with subsequent processing to segment the documents into meaningful sections.

Document metadata was extracted and indexed to facilitate cross-referencing with the Excel database entries. This approach enabled the system to provide more comprehensive responses by supplementing database answers with relevant sections from regulatory documents. The following documents constitute the primary source materials utilized for metadata extraction and analysis.

- FDA: Study Data Standards Resources
- FDA: Specifications for File Format Types Using eCTD Specifications
- PMDA: New Drug Review with Electronic Data
- NMPA: 国家药监局药审中心关于发布《药物临床试验数据递交指导原则（试行）》的通告（2020 年第 16 号）
- NMPA: 申报资料电子光盘技术要求
- NMPA: eCTD 技术规范

DATA PREPROCESSING

Both the Excel database content and the extracted document text underwent several preprocessing steps:

- Text normalization including lowercase conversion and punctuation standardization
- Tokenization using NLTK's word_tokenize function
- Stop word removal to eliminate non-informative terms
- Lemmatization using WordNetLemmatizer to reduce words to their base forms
- Domain-specific term preservation to maintain pharmaceutical terminology integrity

SYSTEM ARCHITECTURE

NLP COMPONENTS

The system implements several NLP techniques to process user queries and match them with appropriate responses:

- Text vectorization using TF-IDF (Term Frequency-Inverse Document Frequency) to convert text into numerical representations
- Word embeddings using Word2Vec to capture semantic relationships between terms
- Cosine similarity calculations to measure the relevance between queries and potential answers
- Named entity recognition using a custom-trained model to identify pharmaceutical and regulatory terminology

MACHINE LEARNING MODELS

Several machine learning approaches were evaluated for query classification and response selection:

A Support Vector Machine (SVM) classifier was implemented for categorizing queries into predefined topics

Random Forest algorithm was used for confidence scoring of potential answers

K-Nearest Neighbors (KNN) approach for identifying similar questions in the knowledge base

Model training was performed using 80% of the database entries, with the remaining 20% reserved for validation. Hyperparameter optimization was conducted using grid search with 5-fold cross-validation to maximize F1-score.

KNOWLEDGE BASE INTEGRATION

The system implements a two-tier search approach:

1. Primary search within the Excel database to find matching question-answer pairs.
2. Secondary search within the document metadata when database matches fall below confidence thresholds. The document metadata encompasses standardized regulations established by CIDSC and various regulatory agencies, including FDA/ NMPA/ PMDA and EMA.
3. Results from both sources are ranked and merged using a weighted scoring algorithm

IMPLEMENTATION

SYSTEM DEVELOPMENT

The question-answering system was developed using Python 3.8 with the following key dependencies:

- pandas (1.1.4) for Excel database manipulation
- scikit-learn (1.3.2) for machine learning algorithms and TF-IDF vectorization
- NLTK (3.6.2) for text preprocessing and linguistic analysis
- PyPDF2 (1.26.0) for PDF text extraction
- gensim (4.0.1) for Word2Vec implementation

The system architecture follows a modular design with separate components for:

- Query preprocessing and normalization
- Knowledge base access and retrieval
- Response ranking and selection
- Output formatting and delivery

QUERY PROCESSING WORKFLOW

When a user submits a question, the system follows this workflow:

1. **Input Processing**:
    - The query undergoes the same preprocessing steps applied to the knowledge base
    - Key entities and terms are identified and weighted
    - Query expansion is performed using domain-specific synonyms
2. **Matching Algorithm**:
    - The processed query is vectorized using the same TF-IDF model applied to the knowledge base
    - Cosine similarity is calculated between the query vector and all question vectors in the database
    - The top 5 matches exceeding a threshold of 0.65 similarity are selected as candidates
    - If no matches meet the threshold, the secondary search in document metadata is triggered
3. **Response Selection**:
    - Candidate responses are ranked based on a composite score combining:
        - Similarity score (60% weight)
        - Category relevance (25% weight)
        - Keyword matching (15% weight)
    - The highest-ranked response is selected for delivery to the user

RESPONSE GENERATION

The system generates responses through the following process:

1. For direct matches from the Excel database, the corresponding answer text is retrieved and formatted
2. For document-based matches, relevant sections are extracted and summarized
3. When appropriate, responses include references to the source documents
4. A confidence score is calculated and included with each response
5. For low-confidence responses, the system indicates uncertainty and suggests alternative interpretations of the query

USER INTERFACE AND INTERACTION

The current implementation provides a simple command-line interface for query input and response display. The system accepts natural language questions and returns formatted responses with source citations when applicable. Future development plans include a web-based interface with additional features such as:

- Query history tracking
- Response feedback collection
- Interactive clarification requests for ambiguous queries

CONCLUSION

This Python-based chatbot demonstrates significant potential in assisting pharmaceutical professionals with regulatory submissions. Its accuracy and speed make it a valuable tool for improving compliance

efficiency. Further enhancements could expand its applicability across global regulatory frameworks. The chatbot can greatly improve our work efficiency with reliable, instant responses.

Future enhancements:

- Integrate live FDA/EMA/PMDA/NMPA portal APIs for real-time updates.
- Implement reinforcement learning for ambiguous query resolution.
- Developing a more sophisticated user interface with visualization capabilities

The system represents a promising step toward more efficient access to regulatory information in pharmaceutical settings, with potential applications in training, compliance, and submission preparation contexts.

REFERENCES

1. FDA. “Study Data Standards Resources” Available at Study Data Standards Resources | FDA
2. FDA. “Specifications for File Format Types Using eCTD Specifications” Available at Specification for Transmitting Electronic Submissions using eCTD Backbone Specifications
3. PMDA. “New Drug Review with Electronic Data” Available at New Drug Review with Electronic Data | Pharmaceuticals and Medical Devices Agency
4. NMPA. “国家药监局药审中心关于发布《药物临床试验数据递交指导原则（试行）》的通告（2020 年第 16 号）” Available at 国家药监局药审中心关于发布《药物临床试验数据递交指导原则（试行）》的通告（2020 年第 16 号）
5. NMPA. “国家药监局药审中心关于更新《申报资料电子光盘技术要求》等文件的通知” Available at 国家药监局药审中心关于更新《申报资料电子光盘技术要求》等文件的通知
6. NMPA. “eCTD 技术规范” Available at eCTD 技术规范 V1.0
7. Chen, L., Wang, H., & Smith, J. (2022). Classification of regulatory queries in pharmaceutical contexts. Journal of Regulatory Science, 10(2), 45-58.
8. Feng, X., Zhang, Y., & Li, Q. (2020). Deep learning approaches for medical consultation systems. IEEE Journal of Biomedical and Health Informatics, 24(6), 1632-1642.
9. Lopez, M., Garcia, R., & Thompson, K. (2019). Automated regulatory information retrieval in financial services. Journal of Information Science, 45(3), 334-349.
10. Zhang, P., & Johnson, T. (2021). Natural language processing for regulatory document analysis in pharmaceutical industry. Artificial Intelligence in Medicine, 112, 102008.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Linjie Jia
Sanofi Aventis
jialinjie@live.cn