A QA System/Chatbot for Info Tunnel

Team 14
Saivarun Tanjore Raghavendra
Tareq Sajjad Ali
Suraj Poldas
Mano Hasha

Dr. Liao Lindi

Introduction to NLP (AIT-526) Sec:001

George Mason University

Introduction:

This project will create a QA system or chatbot in such a way that it would help in answering users' questions by leveraging natural language processing technologies. The system will provide automated answers to queries based on predefined sets of data and, hence, finds broad applicability in number of fields, especially in tunnel NDE techniques. The chatbot shall retrieve information from big pools relating to tunnel technologies but also offers fast and precise answers to specific technical questions.

The system scope will be limited to 26 NDE techniques applied for the evaluation of tunnels and will involve techniques such as AE- Acoustic Emission, GPR-Ground Penetrating Radar, ER- Electrical Resistivity, among others. The chatbot would, therefore, be programmed so that the serving engineers and researchers have access to an instant description of detailed application and methodology of such techniques.

Problem Statement:

The problem it will solve: For engineers and researchers, it's necessary to get quick access to information about different tunnel inspection techniques. As a rule, this information is spread over various documents and in different formats, which is a bit inconvenient for the user. The chatbot unifies this process in that it will allow the users to ask the questions using natural language and get structured and precise answers from an already-prepared dataset.

The challenge is to ensure that such a chatbot would also understand complex queries, map them to appropriate NDE techniques, and furnish information with precision. The solution should address not only user queries but also handle ambiguity and provide responses that are contextually appropriate.

Contributions:

The major contributions of this project are:

- Presenting a complete QA chatbot for NDE techniques.
- There is a dataset integrated which contains 26 various techniques employed in the inspection of tunnels, together with the description in great details.
- Application of the methodology NLP to improve the interaction with the user and understand queries better.
- Real-world data to demonstrate practical usefulness of the system in professional settings.

Related Work:

Various implementations related to NLP-based QA systems and chatbots have been performed. IBM Watson's AI-powered question-answer system provides solutions over various topics in many industries based on big data and machine learning. However, no such system is developed for tunnel NDE techniques. Most of the chatbot implementations are done on general queries or certain business verticals, such as customer service-related or e-commerce-related systems.

The proposed solution is distinct from existing NLP-based QA systems such as IBM Watson in that it has specifically been designed for tunnel Non-Destructive Evaluation (NDE) techniques. While IBM Watson and similar systems respond to broad questions developed utilizing big data interventions and machine learning across many industries, this program focuses on a niche domain, fielding complex engineering queries that may be nonetheless untouchable for general systems. Like the AskMSR question-answering system, which is in simplicity and efficiency by exploiting data redundancy rather than complex linguistic analysis (Brill et al., 2002), this solution uses a targeted approach to dramatically increase performance in the technical field.

Objectives:

The major objectives of the work are as follows:

- To create a chatbot that answers questions concerning the NDE techniques of tunnels.
- The system will have to be able to handle intricate queries requiring multiple layers of information.
- Real-time responses on the basis of the dataset developed that contains information on 26 NDE techniques.
- Technical information in an easy way for engineers, researchers and user working in the field of tunnel inspection.

Dataset Description:

The dataset covers, in great detail, 26 NDE techniques applied in tunnel inspections. Further, each technique is categorized based on functionality, applications, and principles of operation. Included among the methodologies listed in the dataset are AE, GPR, ER, and many others. Each is discussed in a tabulated form that covers descriptions of the given method, working principles, and practical applications.

The analysis in this work could be based on some of the key columns/features listed below:

- Technique Name: Name of particularly NDE method
- Description: Small description of what the method entails
- Applications: Which methods are usually applied and where they are applied
- Working Principle: What scientific basis each method has

Selection of respective features will train a chatbot for the realization of relevance of a piece of information according to particular to users' queries.

Methods and Steps of Data Preprocessing:

1. Data Collection:

Scraping vs. Manual Extraction: After identifying the right sources of data for the QA system, data was retrieved from websites using BeautifulSoup or Scrapy web scraping tools, or the information was manually extracted from the available PDF documents.It is very important during the extraction to obtain the most appropriate and useful data.Regular expressions might be used here to filter to extract particular patterns from the raw data, like keywords or phrases sharing certain technical terminology, which aids in efficiently extracting the right information**.**

2. Data Cleaning:

Pruning out the Non-relevant Information: After collection, removing irrelevant data, duplicate records, or other outliers hindering the system's accuracy is really important. This stage checks the dataset for inconsistencies, such as missing values or duplicates. Regular expressions allow specific patterns of unwanted data to be labeled and cleaned in some cases, for example, poorly formatted text or common spelling errors. In addition, Part of Speech tagging would be needed to ensure that cleaned data has some logical grammatical structures so that the data could be understood properly by downstream natural language processing models. It is possible to complete these tasks using libraries like Pandas, which allow you to deduplicate, impute missing values, and generally make cleaning of datasets more or less straightforward.

3. Feature extraction:

Extraction of Structured Data: Features to extract will be based tightly on what is really necessary from the dataset for that specific purpose, such as columns like Hardcore or others presenting crucial information on certain issues; these features may include the technique name, highly detailed explanation, and application information. Here's the role of POS tagging-in facilitating the identification of technical terms by flagging for key nouns, verbs, and entities within the unstructured text. This guarantees that features are correctly identified and extracted.

Refinement using Regular Expressions: Furthermore, regular expressions would provide additional refinements for the extraction process, allowing specific technical terms or concepts to be singled out. Finally, since technical terms have different usages depending on their context, word sense disambiguation (WSD) techniques shall be applied in order to confirm the relevant meaning of a term.

4. Integration to NLP techniques:

The features are to be integrated with the system once they're extracted, which is where libraries like SpaCy or Hugging Face Transformers facilitate user queries. This allows an easy understanding of the users' intent, since import uses word positions in a user query to identify its structural meaning. Therefore, POS tagging will mark nouns, verbs, and the remaining parts of speech, while giving space for inputs relevant to queries.

Word Sense Disambiguation: WSD does very well in addressing the ambiguities present in user queries, particularly in technical arenas whereby a word carries different meanings based on the context of the user base. This clarifies the need for the chatbot to interpret complex or ambiguous queries accurately.

5. Testing and Validation:

Incremental Testing: The system should be tested incrementally, starting with basic queries and moving to more complex ones. Regular expressions can be used to simulate various types of user queries, including those with ambiguous or incomplete data, to ensure the system’s robustness.

Evaluation Metrics: Use evaluation metrics like Precision, Recall, F1 Score, and User Satisfaction Rate to assess the system’s performance. Additionally, conduct tests to validate how well the system handles word sense disambiguation, ensuring that the chatbot provides accurate answers even when faced with polysemous words or phrases that have multiple meanings.

References

1. Federal Highway Administration. (2021). "Tunnel Non-Destructive Evaluation Technologies." FHWA InfoTechnology. Available at: https://infotechnology.fhwa.dot.gov/tunnel/

2. Gucunski, N., Romero, F., Kruschwitz, S., Feldmann, R., & Parvardeh, H. (2011). "Comprehensive Bridge Deck Deterioration Mapping of Nine Bridges by Nondestructive Evaluation Technologies." Iowa Department of Transportation.

3. Auld, B.A. & Moulder, J.C. (1999). "Review of Advances in Quantitative Eddy Current Nondestructive Evaluation." Journal of Nondestructive Evaluation, 18(1), pp. 3-36. Springer, Basingstoke, England.

4. Huang, S., & Wang, S. (2016). "New Technologies in Electromagnetic Non-destructive Testing." Springer, Basingstoke, England.

5. Langenberg, K.J., Brandfaß, M., Hannemann, R., Kaczorowski, T., & Hofmann, C. (2001). "Inverse Scattering with Acoustic, Electromagnetic, and Elastic Waves as Applied in Nondestructive Evaluation." Wavefield Inversion. Springer, Basingstoke, England.

6. Brill, E., Dumais, S., & Banko, M. (2002). *An analysis of the AskMSR question-answering system*. Microsoft Research.