CS 458/535 - Natural Language Processing

Automatic Questions Tagging System

Abuzar Zulfiqar
19P-0062

February 19, 2023

1 Problem Statement

Sites like Quora and Stackoverflow, which are created specifically to have questions and answers for its users, frequently ask their users to provide five words along with the question so that it may be readily categorized. However, occasionally people give incorrect tags, making it challenging for other users to search through. They need a system that can automatically identify the right and pertinent tags for a user-submitted inquiry in order to fulfil this need.

2 Motivation

For a number of reasons, the issue of manual question tagging is intriguing. First off, manually categorising a huge number of questions might take a lot of time and effort, especially when there are thousands of questions to be sorted. This can result in inaccurate categorization and mistakes, which would be bad for the user experience.
Second, the necessity for effective and efficient organisation of user-generated material is becoming more and more crucial as its volume keeps expanding across numerous platforms. This need can be met by an automatic question tagging system, which offers a quick and precise method of classifying and arranging massive amounts of user-generated questions.

2.1 Example

Let's use an example from a customer care forum to demonstrate this issue. Consider a business that offers a forum for customer help where clients can post inquiries and receive responses from either the support team or other clients. Users may find it challenging to locate the information they require due to the forum's expansion and thousands of inquiries. The business decides to create an automatic question-tagging system that can classify each question according to its topic in order to resolve this problem. For instance, the automatic tagging system may apply tags like "password," "account security," and "login difficulties" to a customer's inquiry on how to change their password, making it simpler for other users to locate similar information in the future.

2.2 Benefits

To increase the effectiveness and efficiency of question-and-answer systems, such as online forums, knowledge bases, and customer service portals, an automatic questions tagging system has been developed. Here are some specific benefits of using an automatic questions tagging system:

- Enhanced searchability: Users can more easily locate the information they're looking for when questions are tagged with pertinent keywords since search filters allow users to focus their results.
- Faster response times: Instead of manually reviewing and classifying each query, human moderators or customer service representatives can save time and effort by using automated tagging.
- Improved data analysis: By tagging questions, you can find trends and patterns in the questions that are asked, which can help you make better decisions and improve your products and services.
- Tagging can assist make sure that questions are sent to the right subject matter experts or knowledge articles, resulting in more accurate and beneficial responses.
- Better user experience: When the system can quickly and precisely identify the material that best fits users' requirements, users can get better support and answers to their inquiries.

Overall, by increasing the effectiveness and efficiency of question and answer systems, an automatic questions tagging system can offer considerable benefits for both users and companies.

3 Background

Readers need have a fundamental understanding of natural language processing (NLP) and machine learning to appreciate the issue of autonomous question tagging. Machine learning is a subset of artificial intelligence (AI) that enables computers to learn from data and improve their performance without being explicitly programmed. NLP is a field of computer science and artificial intelligence that focuses on the interaction between computers and human language.

The idea of tags or labels, which are used to categorise or classify data, should also be recognisable to readers. Tags are used to categorise questions based on their content in the context of automatic question tagging, making it simpler for users to search for and identify pertinent information.

Readers should also be aware of the difficulties of automatically tagging questions. These difficulties include the necessity for a sizable dataset of labelled questions to train and evaluate the system, variances in language use, and ambiguity in meaning.

For a general comprehension of the issue and potential solutions in this sector, readers need have a working knowledge of NLP, machine learning, tags, and the difficulties involved with autonomous question tagging.

4 Related Work

Automatic question tagging systems have been the subject of extensive research, and a number of methods and procedures have been created and put to the test.

One popular strategy is to analyse question text using machine learning techniques and tag the answers appropriately. Decision trees, support vector machines (SVMs), and neural networks are just a few examples of the machine learning algorithms that have been applied in this situation. Based on the properties of new questions, these algorithms can anticipate their tags after being trained on a large dataset of labelled questions [1] .

Another strategy is to employ rule-based systems, which tag questions according to a set of predetermined rules. These guidelines may be based on the text's linguistic or semantic characteristics, such as the existence of particular words or phrases .

Also, some researchers have looked into the usage of hybrid systems, which mix rule-based and machine learning techniques, to increase the precision and effectiveness of automatic question tagging.

Other works being done locally [2] .

Overall, a lot of research has been done in the field of automatic question tagging, and a number of methods and strategies have been created and put to the test. To increase the precision and effectiveness of automatic question tagging systems and to overcome the difficulties involved in this work, research is still being done in this area.

5 Proposed Work

An automatic question tagging system typically involves the following steps:

- Data Collection: A large dataset of labeled questions is required to train and test the system. The dataset may be collected from various sources, such as question and answer websites, customer support forums, or educational platforms.
- Preprocessing: The dataset is preprocessed to clean and normalize the text, remove stop words, and extract features from the text, such as n-grams or part-of-speech tags.
- Training: The preprocessed dataset is used to train a machine learning model, such as a decision tree, SVM, or neural network, to predict tags for new questions. The model is optimized using techniques such as cross-validation and hyperparameter tuning.
- Tagging: The trained model is used to automatically assign tags to new questions. The system may use a rule-based approach in conjunction with the machine learning model to improve accuracy.
- Evaluation: The performance of the system is evaluated using metrics such as precision, recall, and F1 score. The system is refined and retrained as necessary to improve its accuracy and efficiency.

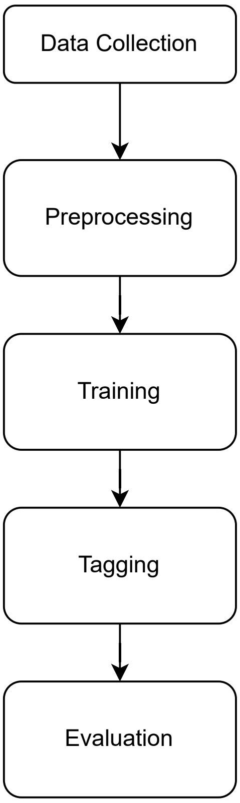

Figure 1: Flow Diagram

6 Evaluation Methodology

To ensure that an autonomous question tagging system is precise and efficient in its duty, evaluation is a crucial step in its development. Precision, recall, and F1 score are just a few of the evaluation measures that can be used to evaluate the system's performance.

How many of the system's tags are accurate is a measure of precision. It is calculated as the system's total number of tags divided by the number of proper tags assigned by the system.

Recall is a gauge of how many of the system's suggested tags are accurate for a particular inquiry. It is determined by dividing the system's assigned correct tags by the total number of accurate tags for the query.

F1 score is the harmonic mean of precision and recall, and is often used as a combined measure of the system's performance.

In addition to these indicators, it's crucial to take into account the automatic question tagging system's particular application and assess how well it works there. For instance, in a customer support forum, comparing the tags that have been assigned to client queries to the actual issues or themes of the customer's query can be used to gauge how accurately the system tags customer questions.

Overall, the evaluation of an automatic question tagging system should be comprehensive and consider the metrics and factors relevant to its specific application.

7 Hypothesis

The goal is to demonstrate that queries may be automatically tagged with pertinent terms in an accurate and timely manner, and that doing so can make it easier for users to access the information they need. By automating the process of classifying and sorting questions, the study may also aim to show how the automatic tagging system might lessen the workload of human moderators or support staff.
The quality of the dataset, the performance of the rule-based system or machine learning model, and the applicability of the assessment measures will all have an impact on how well the study or experiment turns out in the end. If the study yields positive findings, more effective and practical automatic question tagging systems might be created, which would be advantageous for a variety of businesses and applications.

References

[1] Pushpa M Patil, RP Bhavsar, and BV Pawar. A review on natural language processing based automatic question generation. In *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, pages 01–06. IEEE, 2022.

[2] Mihir Prajapati, Mitul Nakrani, Tarjni Vyas, Lata Gohil, Shivani Desai, and Sheshang Degadwala. Automatic question tagging using machine learning and deep learning algorithms. In *2022 6th International Conference on Electronics, Communication and Aerospace Technology*, pages 932–938. IEEE, 2022.