Chapter 2

Question Classification Approaches

2.1 Introduction

The main work which is performed by the classifier is to assign a class for a question depending on classification strategy. For example "Where was Krishna born?" So here what is desired, our question classifier should be like that it will assign this question a category of LOCATION: CITY because the answer we want is a city belonging to location category. So indirectly by guessing the category of the answer, the classification of question is performed. Question classification is also referred as answer style calculation. From this research work set of predefined categories has taken which are considered as question classes usually called question taxonomy or answer type taxonomy. This chapter discusses some basic concepts of question classification.

2.2 Question Classification

For questions classification only two approaches are used which are: rule-based and learning based. There are also some hybrid approaches which combine rule-based and learning based approaches [48], [35], and [49].

Rule based approaches try to match the question with some manually handcrafted rules [50], [37]. These approaches, however, suffer from the need to define too many rules [12]. Moreover, although rule-based approaches could perform well on a particular data set, they may have rather a pitiable performance on a new dataset and also it is difficult to scale them. An example is provided here, which shows the difficulty of rule-based approaches [12]. The following examples belong to the same question category, but they have been formulated in different syntactical forms:

- What is India’s capital city?
- Name the capital of India?
- What's the name of the capital of India?
- Which city is the capital of India?

A simple question can be formulated in many different ways, and making it a very difficult task to manually build specific rules for each different formulation. Alternatively, moving to learning based classifier by means that, training the classifiers through predefined question categories by extracting some features from the question. There are so many approaches which are successfully using the learning based classifiers.

There are several learning that utilizes and demonstrated the learning techniques jointly for rule-based and learning based. The learning which is one of the mainly flourishing mechanism on question classification, initially go through the question with few pre-defined conventions and then utilizes the corresponding convention as features for the learning-based classifier [49]. The similar mechanism is also employed as an effort by Huang et al. [48]. The most successful work which has been done in the past few years regarding question classification are based on the combination of learning-based method and the hybrid method.

### 2.3 Organization of Question Category

The term used for the collection of different classes or categories of question is question taxonomy or question ontology. The various kinds of question taxonomies have been projected in various mechanisms, but the majority of the current learning is supporting the two layer taxonomy which is proposed by Li and Roth [51]. This categorization of question is based on the category of 6 coarse-grained classes and 50 fine-grained classes and it is listed in appendix A. In addition, supplementary question classes which are also well known question taxonomies used for question classification proposed by Hermjakob consisting of 180 classes which is the largest question taxonomy suggested till now [52].
The majority of the current learning-based and hybrid approaches utilizes the taxonomy proposed by Li and Roth the reason is that the author also makes available important set of 6000 marked questions [53]. The above mentioned data set containing two sets of questions one set is for training purpose and another set is for testing purposes, so 5500 questions are in the training set and for testing the accuracy of system 500 questions are provided by the author. This dataset is known as TREC 10 dataset or TREC (Text Retrieval Conference) dataset which is earliest available in University of Illinois Urbana-Champaign (TREC 10).

The improvement in TREC 10 taxonomy with two additional modules or categories that are list and yes-no-explain is provided by Metzler and Croft. Actually, they have produced a distinct data set of 250 questions composed of the MadSci2 question library. MadSci is a technical website which offers a construction in which clients can put a technical question and get an answer from a professional [54].

2.4 Performance Metrics in Question Classification

For measuring the exactness or accuracy of any classifier, it is necessary to calculate its performance on the test set and it can be measured by the formula.

\[
\text{Accuracy} = \frac{\text{correctly classified}}{\text{total}}
\]

The two different measurement metrics regarding classification are: precision and recall they can be employed in question classification problem. The precision and recall can be calculated as:

\[
\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}
\]

\[
\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}
\]

Where cl is any particular class

For the organization in which a question can simply have a single category, a question is properly categorized if the calculated tag is same as the original tag. But for
the organizations which permit a question to be categorized in extra classes that are more than one. A question is accurately categorized, if one of the calculated tag is identical as the original label. [37], [53].