ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(29), DOI: 10.17485/ijst/2015/v8i29/87466, November 2015* ISSN (Online) : 0974-5645

Question Classification: A Review of State-of-the-Art Algorithms and Approaches

S. Jayalakshmi[1*] and Ananthi Sheshasaayee[2]

[1]Periyar University, Salem – 636011, Tamil Nadu, India; jayalakshmi.research@gmail.com
[2]PG and Research Department of Computer Science, Quaid-e-Millath Government College for Women (Autonomous), Chennai – 600002, Tamil Nadu, India; ananthi.research@gmail.com

Abstract

In the recent year Question Answering System achieved significant progress in which Question Classification is an essential component. To design a question classification that assign a semantic category to a given question that represents the type of answer required resulting in an accurate Answer. In this paper present various question classification approaches and techniques, Most of the researches are used to combine Syntactic, Lexical and Semantic features are used to reduce the misclassification. These three techniques are used to improve the efficiency of the question classifications. The existing classifier lacks in accuracy, especially in fine class classification, from the various existing researches are practiced Machine Learning is the best methodology to evaluate the question classification in an appropriate way.

Keywords: Machine Learning, Question Classification, Semantic Similarity, Syntactic Features

1. Introduction

Question answering systems have inherited many techniques from machine learning, information retrieval, and natural language processing over the years to retrieve precise answers automatically. The QC has a significant role in the question answering system. The syntactic and semantic structure representation of QC and question answering system are described in[1]. Classification is to label a question into a class that represents the answer type. Questions can be widely categorized as factoid and non-factoid questions. If the information in a question is a simple fact, it is a factoid question. The answer type of factoid question has been just one or more words that correspond to a named entity such as a person, organization or location. The answer type of non-factoid question is complex or descriptive. The QC is a vital part of the question answering system as the selected question type can be used for various purposes. It is a component of other potential applications associated to information retrieval and Natural Language Processing (NLP). Machine learning approach efficiently manages automatic classification of text documents by learning from a set of labeled training documents[2]. A state of the art survey of QC presents different approaches and newly emerging methods[3].

The process of question classification helps to find the answer type of the question. They have different approaches in the question classification, namely rule-based approach and machine learning approaches[4]. Question analysis is the first stage of the question answering system. IBM Watson system performs detailed analysis of questions to assist the answering system to deliver the accurate results[5]. The questions are classified into fine-grained classes using a hierarchical classifier. The classifier learns from a layered semantical hierarchy of answer types[6]. Flat taxonomies have only one level of class without having sub-classes while hierarchical taxonomies have multi-level classes. Statistical learning methods are the successful approaches for classifying the questions[7]. Furthermore, these methods extract lexical, syntactic and semantic features of questions. Now-a-days most of the question classifier depends on supervised machine learning approach. This work combines with lexical, semantic and syntactical features to improve the accuracy of classification. The question type of taxonomy is a set of predefined categories that considered a question classes. The question type of taxonomy consists of six coarse-

* *Author for correspondence*

grained classifiers such as abbreviation, description, entity, human, location and numeric, and a set of fine-grained classifiers.

2. Natural Language Processing (NLP), Semantic and Syntactic Analysis

2.1 Natural Language Processing

NLP is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things.

2.2 Semantic Analysis

Semantic feature is a semantically related word that associated with an exact question class. Semantic features include the Word Net relation, Hypernyms, named entity, antonyms, synonyms and semantic headwords. Word Net is a lexical database of words that gives a lexical hierarchy that relates a word with a high level of semantic concepts in terms of synonyms and antonyms.

2.3 Syntactic Analysis

It represents syntax related features that parse the structure of the question based on the grammar. Syntactic features include question head-words and Part-Of-Speech (POS) tags. The headword is the most informative word and a parser extracts headword of the question. The parse tree represents a syntactic structure of a sentence based on grammar rules. The parse tree of a question extracts the pre-terminal nodes to use as features. These nodes denote the part-of-speech (POS) tags of the question tokens. The POS tag of a question contains words such as Noun (N), Noun Phrase (NP), Verb Phrase (VP), and adjective (JJ).

3. Question Classification Approaches

3.1 Rule-based Question Classification

The rule-based classifier classifies the question in straightforward method using a set of predefined heuristic rules based on taxonomy. The rule-based approach classifies the question using manually crafted rules by experts. A rule based classifier is presented and evaluated in[8]. It uses rules to detect the question headword and utilize Word Net to map the target category. In this method, machine learning classifier uses the features of a rule based classifier to derive the final category. A novel and unified paradigm based on a rule based method and statistical method is proposed in[9]. The proposed QC is designed with a Markov logic network, which uses a fuzzy discriminative learning approach. The rule based approach does not support other domains or different language as it is difficult to frame a new set of rules. The rule-based approaches perform well in a particular dataset and rather than the reduced performance of the new dataset. The rule based approach is accurate in predicting a certain category of questions; however, it is not scalable to a large number of questions and syntactical structures.

3.2 Machine learning Question Classification

Machine learning is a type of artificial intelligence that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change. Machine Learning (ML) is coming into its own, with a growing recognition that ML can play a key role in a wide range of critical applications, such as data mining, natural language processing, image recognition and expert systems. ML provides potential solutions in all these domains and more. It is divided into two broad categories are.

3.2.1 Supervised Machine Learning

The program is "trained" on a pre-defined set of "training examples", which then facilitate its ability to reach an accurate conclusion when given new date. Supervised learning is fairly common in classification problems. Supervised learning is the most common technique for training neural networks and decision trees. Both of these techniques are highly dependent on the information given by the pre-determined classifications. In the case of neural networks, the classification is used to determine the error of the network and then adjust the network to minimize it, and in decision trees, the classifications are used to determine what attributes provide the most information that can be used to solve the classification puzzle.

3.2.2 Unsupervised Machine Learning

Unsupervised learning seems much harder: the goal is to have the computer learn how to do something that we

tell it how to do! There are actually two approaches to unsupervised learning.

The first approach is to teach the agent not by giving explicit categorizations, but by using some sort of reward system to indicate success. Note that this type of training will generally fit into the decision problem framework because the goal is not to produce a classification but to make decisions that maximize rewards.

A second type of unsupervised learning is called clustering. In this type of learning, the goal is not to maximize a utility function, but simply to find similarities in the training data. The assumption is often that the clusters discovered will match reasonably well with an intuitive classification. Clustering individuals based on demographics might result in a clustering of the wealthy in one group and the poor in another group.

ML techniques based QC uses lexical, syntactic and semantic features. An Extreme Learning Machine (ELM) uses semantic features to improve both training and testing compared to the benchmark of the SVM classifier[10]. The function of ELM is to classify the semantic features of statistical QC.

3.3 Hybrid Approach

The hybrid approach combines the feature set of two or more QC techniques. Hybrid approaches combines the concept of rule-based and learning based technique, it proposed the hybrid approach that exploits the information of headwords and categories from rule-based classifier to generate feature set for training and merges this information with the information obtained from question unigrams[11].

Question topic classification is a major problem of Community Question Answering (CQA) services[12]. It is difficult to learn the performance of the newly emerging methods in QC or short text classification. It combines Naïve Bayes (NB), Support Vector Machine (SVM), and Maximum Entropy (ME) approaches using different feature representation in some component of QC.

4. Other Question Classification Systems

QASYO is a unified question answer framework, exploits natural language processing, YAGO ontology and information retrieval techniques[13]. It accepts questions in natural language and maps the query into the target ontology knowledge base. In question analysis, it uses a natural language query to attain a formal representation in the direction of compliant through the YAGO query formulation database. The classifier considers the QC as a tree classification problem[14]. It suggested a sub tree training for tree classification and utilizes maximum entropy and a boosting model as a question classifier. The powerful of Word Net semantic features and in addition to the Wikipedia knowledge repository guides informative terms explicitly to process the questions[15].

5. Future Directions

The next generation of the QC system needs to understand more complex questions and different forms of questions. In order to categorize questions, effective NLP and machine learning are necessary to meet the future requirements of QC. Robust lexical, syntactic, and semantic parsers are absolutely necessary for understanding complex question patterns. A new class of hybrid machine learning classifier is the definite future requirement.

6. Conclusion

The Question Classification is a crucial component to retrieve an effective answer for the posted query. In this paper present various question Classification approaches and its features, the researchers are analyzed various approaches from their view point the Machine Learning Approach is suggested to improve the fine class classification. It produces better result than the Rule based approach. By enhancing the features it may produce some complex and noisy information it leading to mis-classification. In order to train the Learning the Syntactic and Semantic features are used to improve the efficiency of the QC accuracy. In order to train the learning algorithm, it can be used as an effective set of lexical, semantic and syntactic features.

7. References

1. Brank J, Malden D, Grobelnik M. Large-scale Hierarchical Text Classification using SVM and Coding Matrices. 2010.
2. Garcia Cumberers MA, Urena Lopez LA, Santiago FM. BRUJA: Question classification for Spanish. Using machine translation and an english classifier. 2006.

3. Li X, Roth D. Learning question classifiers**.** 2002.
4. Gharehchopogh FS, Lotfi Y. Machine learning based question classification methods in the question answering systems. Int J Innovat Appl Stud. 2013; 4(2).
5. Silva J, Coheur L, Mendes A, Wichert A. From symbolic to sub symbolic information in question classification. Artif Intell Rev. 2011; 35(2):137–54.
6. Metzler D, Croft WB. Analysis of statistical question classification for fact-based questions. 2004.
7. Hovy E, et al. Question answering in webclopedia. Proc 9th Text REtrieval Conf (TREC-9). 2002. p. 655–64.
8. Hovy E, et al. A question/answer typology with surface text patterns. Proc 2nd Int Conf Human Language Technology Research. HLT'02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 2002. p. 247–51.
9. May R, Steinberg A. Building a question classifier for a TREC-style question answering system. 2004.
10. Li X, Roth D. Learning question classifiers: The role of semantic information. 2004.
11. Hermjakob U. Parsing and question classification for question answering. Appeared in Proceedings of the Workshop on Open-Domain Question Answering at ACL. 2001.
12. Wagstaff KL. Machine learning that matters. Proceedings of the 29th International Conference on Machine Learning California Institute of Technology. 2012.
13. Collins M, Singer Y. Unsupervised models for named entity classification. Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 1999. p.100–10.
14. Clark A. Inducing syntactic categories by context distribution clustering. Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on computational natural language learning. 2000; 7:91–4.
15. Mishra M, Mishra VK, Sharma HR. Question classification using semantic, syntactic and lexical features. International Journal of Web and Semantic Technology. 2013; 4(3).