Information Processing and Management xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Question categorization and classification using grammar based approach

Alaa Mohasseb*, Mohamed Bader-El-Den, Mihaela Cocea

School of Computing, University of Portsmouth, United Kingdom

ARTICLE INFO

Keywords:
Question classification
Machine learning
Text mining
Text classification
Natural language processing (NLP)

ABSTRACT

Question-answering has become one of the most popular information retrieval applications. Despite that most question-answering systems try to improve the user experience and the technology used in finding relevant results, many difficulties are still faced because of the continuous increase in the amount of web content. Questions Classification (QC) plays an important role in question-answering systems, with one of the major tasks in the enhancement of the classification process being the identification of questions types. A broad range of QC approaches has been proposed with the aim of helping to find a solution for the classification problems; most of these are approaches based on bag-of-words or dictionaries. In this research, we present an analysis of the different type of questions based on their grammatical structure. We identify different patterns and use machine learning algorithms to classify them. A framework is proposed for question classification using a grammar-based approach (GQCC) which exploits the structure of the questions. Our findings indicate that using syntactic categories related to different domain-specific types of Common Nouns, Numeral Numbers and Proper Nouns enable the machine learning algorithms to better differentiate between different question types. The paper presents a wide range of experiments the results show that the GQCC using J48 classifier has outperformed other classification methods with 90.1% accuracy.

1. Introduction

Question-answering has become one of the most popular information retrieval applications. Questions Classification (QC) plays an important role in question-answering systems and one of the major tasks in the enhancement of the classification process is the identification of questions types.

Despite that most Question-Answering Systems (QASs) try to improve the user experience and the technology used in finding relevant results, many difficulties are still faced because of the continuous increase in the amount of web content and the low response rate to many questions Liu and Jansen (2017) and Liu and Jansen (2018). The goal of the question classification process is to accurately assign labels to questions based on an expected answer type Metzler and Croft (2005).

The task of generating answers to the users questions is directly related to the type of questions asked Moldovan, Paşca, Harabagiu, and Surdeanu (2003). Hence, the classification of the questions performed in QASs directly affects the answers. Results show that most errors happen due to miss-classification of questions performed in QASs Moldovan et al. (2003). Authors in Bu, Zhu, Hao, and Zhu (2010) performed function oriented classification of questions by integrating pattern matching and machine learning techniques, while Benamara (2004) classify questions by taking account of their expected types of responses. In

* Corresponding author.
E-mail addresses: alaa.mohasseb@port.ac.uk (A. Mohasseb), mohamed.bader@port.ac.uk (M. Bader-El-Den), mihaela.cocea@port.ac.uk (M. Cocea).

https://doi.org/10.1016/j.ipm.2018.05.001
Received 11 December 2017; Received in revised form 30 April 2018; Accepted 10 May 2018

addition, Kolomiyets and Moens (2011) stated that question type is defined as a certain semantic category of questions characterized by some common properties.

Recent studies classified users' questions using different features like bag-of-words Zhang and Lee (2003), Li, Huang, and WU (2005), Yen et al. (2013), Mishra, Mishra, and Sharma (2013), semantic and syntactic features Yen et al. (2013), Hardy and Cheah (2013),Song, Wenyin, Gu, Quan, and Hao (2011), and uni-gram and word shape features Huang, Thint, and Qin (2008). Authors in Huang et al. (2008) stated that features are the key to obtain an accurate question classifier. Furthermore, in order to distinguish between different types of questions, many previous studies classified questions using different machine learning algorithms.

Support Vector Machine (SVM) is one of the most used algorithms Metzler and Croft (2005), Bullington, Endres, and Rahman (2007), Huang et al. (2008), Hao, Xie, and Xu (2015), Van-Tu and Anh-Cuong (2016), Hasan and Zakaria (2016), Xu, Cheng, and Kong (2016). According to authors in Mishra et al. (2013) combining an SVM classifier with semantic, syntactic and lexical features improves the classification accuracy. Other works like Zhang and Lee (2003), Song et al. (2011), Mishra et al. (2013) and Mohd and Hashmy (2018) used SVM in addition to other machine learning algorithms such as Naive Bayes, Nearest Neighbors and Decision Tree. Moreover, works like Sagara and Hagiwara (2014) and Ture and Jojic (2016) used Neural Networks as the machine learning algorithm.

In this study, we propose a new grammar-based framework for questions categorization and classification (GQCC). The GQCC framework represented the question as a grammatical pattern i.e. each term is replaced by its corresponding grammatical category and all grammatical categories in the question form the grammatical pattern. In addition, domain-specific grammatical categories are used as the grammatical categories and are not just the standard English ones. Furthermore, in order to transform the question into a grammatical patterns a formal grammar approach is used and a machine learning is applied on this transformed data to obtain models for automatic classification.

The rest of the paper is organised as follows. Section 2 outlines previous work in question classification, including different question taxonomies, as well as previous classification approaches using machine learning techniques. Section 4 describes the proposed question classification framework. Section 3 highlights the research objectives. The experiments setup and results are presented in Section 5, while the results are discussed in Section 6. Finally, Section 7 concludes the paper and outlines directions for future work.

2. Background

In this section we review previous work on question classification according to user intent. Different categories of user intent are outlined in Section 2.1, while Section 2.2 reviews previous work on question classification methods.

2.1. Questions categories

Different categories of questions were defined, which are summarised in Table 1. According to authors in Kolomiyets and Moens (2011) the major question types are: factoids, list, definition, hypothetical, causal, relationship, procedural, and confirmation questions. A factoid question is a question which usually starts with a Wh-interrogated word (What, When, Where, Who) and requires as an answer a fact expressed in the text body. On the other hand, a list question is a question, which requires as an answer a list of entities or facts; a list question usually starts as: List/Name [me] [all/at least NUMBER/some]. Furthermore, a definition question is a question, which requires finding the definition of the term in the question and normally starts with "What is". Related to the latter is the descriptive question, which asks for definitional information or for the description of an event, and the opinion question whose focus is the opinion about an entity or an event. A hypothetical question is a question, which requires information about a hypothetical event and has the form of "What would happen if". In addition, a causal question is a question which requires explanation of an event or artifact, typically starting with "Why". A relationship question asks about a relation between two entities, while a procedural question is a question which requires as an answer a list of instructions for accomplishing the task mentioned in the question. Finally, a confirmation question is a question, which requires a Yes or No as an answer to an event expressed in the question.

The classification in Bu et al. (2010) was motivated by related work on user goal classification by Broder Broder (2002) and Rose and Levinson Rose and Levinson (2004). The proposed function-based question classification categories were tailored to general QA, containing six types, namely: Fact, List, Reason, Solution, Definition and Navigation. For the Fact type of question the expected answer will be a short phrase; these questions are asked to get a general fact as an answer. For the List type of question each answer

Table 1
Summary of user intent categories for questions .

Authors	Categories
Kolomiyets and Moens (2011)	factoids, list, definition, hypothetical, causal, relationship, procedural, and confirmation questions.
Bu et al. (2010)	Fact, List, Reason, Solution, Definition and Navigation.
Bullington et al. (2007)	Advantage/Disadvantage, Cause and Effect, Comparison, Definition, Example, Explanation, Identification, List, Opinion, Rationale and Significance.
Li and Roth (2006)	Abbreviation, Description, Entity, Human, Location and Numeric as coarse classes; and Expression, Manner, Color, City.