Question Categorization and Classification using Grammar Based Approach

ALAA MOHASSEB[1], MOHAMED BADER-El-DEN[1] and MIHAELA COCEA[1]
[1]School of Computing, University of Portsmouth, United Kingdom
Emails:{alaa.mohasseb, mohamed.bader, mihaela.cocea}@port.ac.uk

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.

Keywords:
Question Classification, Machine Learning, Text Mining, Text Classification, Natural Language Processing (NLP).

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 [27] and [28]. The goal of the question classification process is to accurately assign labels to questions based on an expected answer type [30].

The task of generating answers to the users questions is directly related to the type of questions asked [37]. 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 [37]. Authors in [5] performed function oriented classification of questions by integrating pattern matching and machine learning techniques, while [2] classify questions by taking account of their expected types of responses. In addition, [20] 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 [54], [24], [53], [31], semantic and syntactic features [53], [13],[49], and uni-gram and word shape features [17]. Authors in [17] 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 [30], [6], [17], [12], [51], [14], [52]. According to authors in [31] combining an SVM classifier with semantic, syntactic and lexical features improves the classification accuracy. Other works like [54], [49], [31] and [36] used SVM in addition to other machine learning algorithms such

Table 1: Summary of user intent categories for questions

Authors	Categories
[20]	factoids, list, definition, hypothetical, causal, relationship, procedural, and confirmation questions
[5]	Fact, List, Reason, Solution, Definition and Navigation.
[6]	Advantage/Disadvantage, Cause and Effect, Comparison, Definition, Example, Explanation, Identification, List, Opinion, Rationale and Significance.
[25]	Abbreviation, Description, Entity, Human, Location and Numeric as coarse classes; and Expression, Manner, Color, City.

as Naive Bayes, Nearest Neighbors and Decision Tree. Moreover, works like [47] and [50] 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 [20] 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 [5] was motivated by related work on user goal classification by Broder [4] and Rose and Levinson [46]. 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 will be a single phrase or a phrase with explanations or comments; these questions are asked to get a list of answers. Furthermore, a good answer summary should contain a variety of opinions or comprehensive explanations for Reason Type of question in which Sentence-level summarization can be employed; these questions are asked to get opinions or explanations as the answer. For the Solution type of questions, the sentences in an answer usually have a logical order, thus the summary task cannot be performed on sentence level; these questions are asked to solve a problem. The Definition type of questions are asked to get a description of concepts as an answer; usually this information can be found in Wikipedia. If the answer is too long, it should summarized into a shorter one. Finally, Navigation type of questions are asked to find websites or resources; sometimes the websites are given by name and the resources are given directly.

Authors in [6] classified open-ended questions to 11 categories, which are: Advantage/Disadvantage, Cause and Effect, Comparison, Definition, Example, Explanation, Identification, List, Opinion, Rationale, and Significance. Advantages and disadvantages are questions that may require certain number, while Cause and Effect are questions that explain the effect of something on something else. Moreover, a Comparison question requires and answer that outlines differences and/or similarities between two or more entities. Furthermore, a Definition question requires a relatively short explanation or description (just few lines or few sentences). On the other hand, an Example question requires an answer that provides an example. An Explanation question provides more explanation or more details than the what questions. Identification questions provide answers allowing the identification of something. The List question provides a list of points which may or may not be in sequence. Opinion questions give as answers personal opinions on a particular point or a statement supporting an argument or advocating against it. Finally, the answer to a Rationale question explains why a statement/question is true or false, while an answer to a Significance question explains the importance of something or why it may be important.

The most famous expected answer type taxonomy with regard to factoid questions is the one of Li and Roth [25]. Their two-layer taxonomy consists of a set of six coarse-grained categories and fifty fine-grained ones, e.g., Abbreviation, Description, Entity, Human, Location and Numeric as coarse classes, and Expression, Manner, Color and City as fine-grained classes. This classification deals with factoid questions which is a very limited class of real world questions.

In previous work, many researchers focused on particular types of questions. For example, work in [15] focused on the ”causal” question type, while work in [2, 30, 51] focused on factoid questions.

Most work based on Li and Roth [25] classification of question [24, 30, 51, 52, 21, 29, 54, 31, 17, 39, 38, 23] focus on factoid questions since the categorization proposed by Li and Roth mainly deal with factoid questions.

2.2. *Question Classification Methods*

In this section we review related work about question classification methods and machine learning algorithms.

Many recent studies classified users' question using different features like bag-of-words [54], [24], [53], [31], semantic and syntactic features [53], and uni-gram and word shape features [17]. Furthermore, to distinguish between different types of questions, many previous studies classified questions using different machine learning algorithms.

Authors in [17] proposed head word features, which is one single word specifying the object that the question seeks, and used two approaches to augment the semantic features of such head words using WordNet. In addition, other standard features were augmented, which means some features were increased, such as wh-word, unigram feature, and word shape feature. In [53] a framework has been proposed, which integrates a question classifier with a simple document/passage retriever, and proposed context-ranking models. This method provides flexible features to learners (machine learning algorithms), such as word forms, syntactic features, and semantic word features. In addition, the proposed context-ranking model, which is based on the sequential labeling of tasks, combines rich features like full parsers, predefined syntactic patterns, and more training materials to predict whether the input passage is relevant to the question type.

The work in [24] used machine learning approaches, namely, different classifiers and multiple classifier combination methods by using compositive statistic and rule classifiers, and by introducing a dependency structure from Minipar and linguistic knowledge from Wordnet into question representation. In addition, features like the Dependency Structure, Wordnet Synsets, Bag-of-Words, and Bi-gram were used. Also a number of kernel functions were used and the influence of different ways of combining classifiers, such as Voting, adaboost, Artificial Neural Networks (ANN) and Transition-Based Learning (TBL), on the precision of question classification was analysed.

In [11] a hybrid approach was proposed, named ATICM which is based on dependency tree analysis for automated answer type identification and classification by utilizing both syntactic and semantic analysis. This method contains a compact WordNet-based hypernym expansion strategy to classify identified question target words into question target categories. Result showed that ATICM approach has achieved an accuracy of 93.9% on the UIUC dataset and 92.8% on the TREC10 dataset. In addition, authors in [51] proposed a method of using a feature selection algorithm to determine appropriate features corresponding to different question types. Moreover, they designed a new type of feature, which is based on question patterns; then applied a feature selection algorithm to determine the most appropriate feature set for each type of questions. The proposed approach was tested on the benchmark dataset TREC, using SVM for the classification algorithm.

In [30] a statistical classifier has been proposed which is based on SVM and uses prior knowledge about correlations between question words and types in order to learn question word specific classifiers, i.e. a what question will be classified with SVM*what*. In addition, any data set, question ontology, or set of features can be used with this statistical framework. Furthermore, [52] proposed a SVM-based approach for question classification. In addition, a dependency relations and high-frequency words are incorporated into the baseline system. Experiments on the UIUC corpus showed that the introduced features can improve the baseline system significantly in which the combination of top word and dependency relation features improved the accuracy to 93.4%.

Other works like [54] and [31] used SVM in addition to other machine learning algorithms. [31] proposed an approach for question classification through using three different classifiers, k-Nearest Neighbor (kNN), Nave Bayes (NB), and SVM, using two kinds of features: bag-of-words and bag-of-ngrams. In order to train the learning algorithm, a set of lexical, syntactic, and semantic features were used, among which are the question headword, which is a word in a given question that represents the information that is being sought, and hypernym which is a word with higher level semantic concepts. Similarly, in [54] five machine learning algorithms were used, KNN, NB, Decision Tree (DT), Sparse Network of Winnows (SNoW), and SVM, using two kinds of features: bag-of-words and bag-of-ngrams.

SVM were also used in [6] for the classification of open-ended questions. They have stated that SVM could be trained to recognize the occurrence of certain keywords or phrases in a question class and then, based on the recurrence of these same keywords, be able to correctly identify a question as belonging to that class. Another classification approach has been proposed in [9] using SVM. According to the authors in this work an enormous amount of time is required to create a rich collection of patterns and keywords for a good coverage of questions in an open-domain application, so they have used support vector machines for question classification. The goal is to replace the regular expression based classifier with a classifier that learns from a set of labeled questions and represented the questions as frequency weighted vectors of salient terms. Moreover, works like [47] and [50] used Neural Networks as the machine learning algorithm. [47] proposed a neural network for a question answering system. The proposed network is composed of three layers and one network: Sentence Layer, Knowledge Layer, Deep Case Layer and Dictionary Network. The input sentences are divided into knowledge units and stored in the Knowledge Layer.

In [26] a classification method was proposed for community question answering (CQA) system based on ensemble learning, using supervised learning and semi-supervised learning of different feature extraction methods like lexical semantic extension and different classifiers in the question classification, the supervised learning and the semi-supervised learning adopt three different classifiers, which are J48graft, J48 and Nave Bayes. The experiments verified that the semi-supervised classification algorithm based on ensemble can effectively utilize a mass of unlabeled question samples to enhance the classification accuracy. Finally, the proposed approach in [50] formulates the task as two machine learning problems, which are, detecting the entities in the question, and classifying the question as one of the relation types in the knowledge base. Based on this assumption of the structure, this approach trained two recurrent neural networks and outperformed state-of-the-art approaches by significant margins; the relative improvement reached 16% for Web Questions, and surpassed 38% for Simple Questions.

Unlike the previous approaches, we propose a grammar-based framework for questions categorization and classification which deals with different types of questions and different domain categories by exploiting the structure of the question through using general and domain-specific grammatical categories and rules. Moreover, the grammar provides a flexible and powerful platform for integrating prior-domain information about each question category into the tagging and classification phases. Details of the framework are presented in the next section.

3. Research Objectives

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 pattern a formal grammar approach is used and machine learning is applied on this transformed data to obtain models for automatic classification. The aim of the research presented in this paper is to:

1. Evaluate the influence of using the structure of a question and the domain-specific grammatical categories on the classification performance.
2. Investigate the impact of using different levels of detail of grammatical categories and domain-specific information on the classification performance and compare the classification performance of different machine learning algorithms.

4. Grammar-Based Framework for Question Categorization and Classification

We propose a Grammar Based Framework for Question Categorization and Classification (GQCC), shown in Figure 1. GQCC takes into account the grammatical structure of the questions and combines domain-related information with grammatical rules and patterns. The aim of this approach is to create a question categorization and classification framework that could easily be applied to different question-answering systems by creating domain specific grammatical rules and patterns for each type of question.

GQCC transforms the question using grammatical rules into a new form of representation in which each term in the question is represented as its grammatical category, which we call Grammatical Pattern, which has the advantage of preserving the grammatical structure of the question. The grammatical rules contain in addition to typical categories of English grammar, domain-related grammatical categories. The three phases of the Grammar Based Framework for question categorization and classification (GQCC) are described in the following subsections.

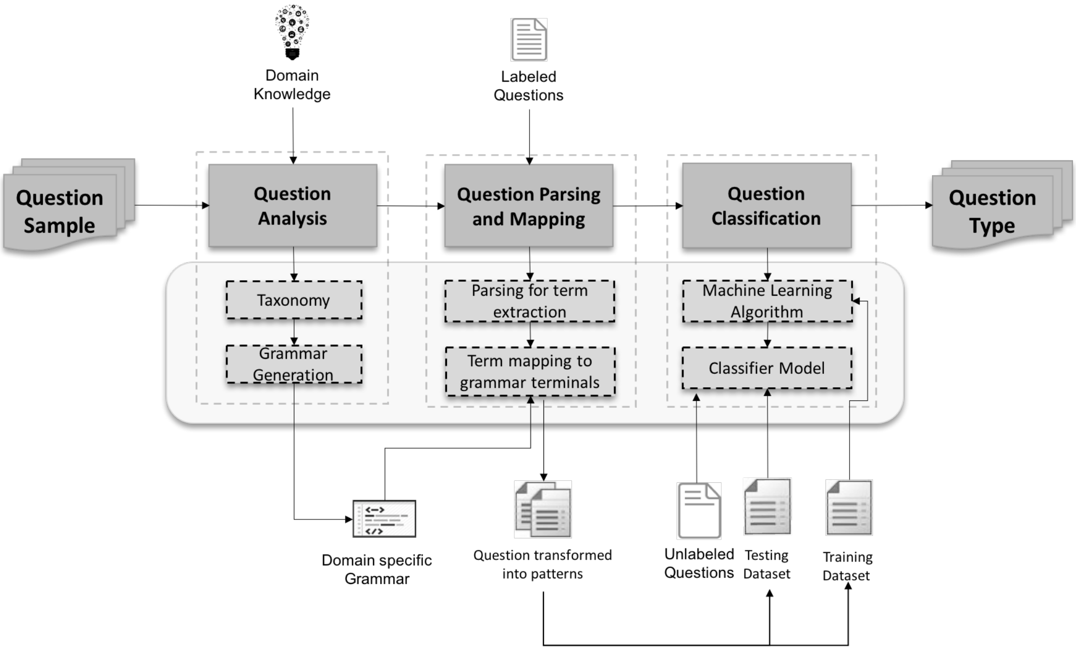


Figure 1: Question Classification Framework

4.1. Phase I: Question Analysis

The proposed GQCC framework makes use of two related sources of information about the questions, i.e. the structure of different questions and question domain-specific information available about each category of questions. In order to capture the relation between these two sources and combine them in a unified structure, a formal grammar is designed for the question classification problem. There are several forms of grammar; the context free grammar in the Backus normal form (BNF) is adopted in this study as it is the most widely used grammar in computing. Eventhough BNF can not provide a full description of the English grammar [19], [42], [40] the target in this research is to use a simple version of the English grammar combined with domain-specific grammatical categories to guide the question classification and categorization stage.

4.1.1. Analysis of Questions Structure and Characteristics

We propose a new question categorization which is based on the general question types. The objective of this classification is to focus on the general and simple type of questions that are asked by most people. This classification is motivated by the basic English grammar [22], [8] and by the categorization of questions in [20],[5], [6]. After the analysis of $10,000$ questions from different datasets, i.e. Yahoo Non-Factoid Question Dataset[1], TREC 2007 Question Answering Data[2] and a Wikipedia dataset[3] that was generated by [48], $5,000$ questions were randomly selected and labelled based on the proposed question categorization. Furthermore, 200 were randomly selected for the validation process and another $1,160$ were selected for the experiments.

Questions were classified into six different categories, which are: causal, choice, confirmation (Yes-No Questions), factoid (Wh-Questions), hypothetical and list. These categories are based on the question types in English and the classification is based on types of questions asked by users and the answers given. Each of these questions has its own characteristics, features, and structure that help in the identification of each type. The choice, confirmation (Yes-No Questions), factoid (Wh-Questions) and hypothetical questions were adapted from the English grammar, while list and causal were adapted from previous works. Table 2 shows a summary of the question types structure and characteristics which are detailed below.

Table 2: Question Types Structure and Characteristics

Questions	**Answer**	**PoS that identify the Question**
Confirmation	Yes or No	AuxV
Factoid	Any kind of information could be given as an answer	QW
Choice	a selection between two or more options	Conj (OR) , LV, AuxV
Hypothetical	Any kind of information could be given and could have more than one accurate answer	QW (What)
Causal	Deep explanations and elaborations related to the topic in the question	QW (Why, How)
List	A list of different Facts, Entities, Events and Names, depend on the topic.	Plural (CN), QW (What, Which, Who)

1. *Yes-No Questions (Confirmation Questions):* This type of question begins with an auxiliary verb or linking verb, and the expected answer is either 'Yes' or 'No', for example *"is Detroit a city in Michigan?"*. In addition, the question could start with negative auxiliary verbs or linking verbs, such as *"Wasn't Leonardo da Vinci born on April 15?"*. Moreover, this type of question usually does not contain a question word.
2. *Wh-Questions (Factoid Questions):* The main feature of this type of question is the presence of question words, e.g. *"What did Alessandro Volta invent in 1800?"*; any kind of information can be given as a response. Furthermore, most of them start with a question word, such as *What / Where / Why / Who / Whose / When / Which*.

[1] https://ciir.cs.umass.edu/downloads/nfL6/
[2] http://trec.nist.gov/data/qa/t2007_qadata.html
[3] https://www.cs.cmu.edu/~ark/QA-data

However, there are other question words that do not start with "wh" as well, e.g. *how / how many / how often / how far / how much / how long / how old*. In addition, the structure of the question could begin with a Preposition followed by a question, *"P + QW"*, rather than a question words, such as *"In what year was Nairobi founded?" / "at what distance does the earth curve?"*. Also in many cases the question word could be found in the middle of the question, for example *"water boils at what temperature?"*. Most factoid questions are formulated as an advice question, e.g. *"how do you quit smoking?"*, and are related to facts, current events ideas and suggestions. In addition, some factoid questions could contain two types of questions, factoid and causal, for example *"what is a good phone service and why?"*.

3. *Choice Questions:* The structure of this type of question mainly offers choices in the question; usually the question contains two (or more) presented options. These options are connected using the conjunction "OR". Questions in this type could begin with a: (a) linking verb, e.g. *"was ancient Egypt before or after ancient Greece?"*; (b) Auxiliary Verb, e.g. *did Einstein die in the 50s or 60s?*; (c) Question word, e.g. *"what is better Samsung or iPhone?"* or (d) Determiner, e.g. *"which is better Netflix or Amazon?"*. Furthermore, some choice questions could contain causal questions, For example, *"Which is better Playstation or Xbox 360 and why?"*
4. *Hypothetical Questions:* a hypothetical question is asked to have a general idea of a certain situation. The question typically begins with the question word "What", e.g. *"what would you do if someone had a heart attack?"*; *"what would happen if the nervous system stopped working"*. It is mainly a what/if question.
5. *Causal Questions:* the structure of this question begins with the question words "How" or "Why" and the answer requires further explanation; for example, *"why do clouds turn dark when it's about to rain?"*. However, the question could begin with *"if"*, and takes the following format *"if...then...why"* or *"if...then...how"*. In addition, causal questions could in many cases begin with a question word followed by a negative linking verb or a negative auxiliary verb, for example *"why isn't my phone connecting to wifi?"*.
6. *List Questions:* The answer of this type of question takes the form of a list of entities or facts. Plural terms are a highly reliable indicator of this question. In addition, this question often begins with the words *"List"* or *"Name"* (e.g. *"list of Disney movies" / "Name of dinosaurs"*) or a question word followed by a plural term, such as *"what countries are in Europe?"*, *"which products contain gluten?"*. However, in some cases list questions could begin with a preposition followed by a question word, for example *"in what countries does Uber operate?"/ "in which African countries is French spoken?"*.

4.1.2. Validation of Questions Types Categories

A validation set was created by having three annotators independently judge 200 questions that were randomly selected from a sample of 5,000 that was obtained from the three data-sets mentioned previously: Yahoo Non-Factoid Question Dataset , TREC 2007 Question Answering Data and a Wikipedia dataset that was generated by [48].

Questions were labelled by assessors according to the categorization of questions that was discussed in Section 4.1.1. In the first stage, two annotators labelled the questions, and then we reviewed the classification results. If a question was labelled differently by the two annotators, a third annotator assigned a label to the question. The two annotators disagreed on 5.5% of the questions.

4.1.3. Terms Taxonomy

The terms taxonomy has been used for the purpose of transforming the questions (by using the grammar) into a new representation as a series of grammatical terms, i.e. a grammatical pattern.

The terms taxonomy is mainly based on the seven major word classes in English, which are Verb (V), Noun (N), Determiner (D), Adjective (Adj), Adverb (Adv), Preposition (P) and Conjunction (Conj). In addition, we added a category for question words (QW) that contains the six main question words: "how", "who", "when", "where", "what" and "which". Some word classes like Noun can have sub-classes, such as Common Nouns (CN), Proper Nouns (PN), Pronouns (Pron) and Numeral Nouns (NN), as well as Verbs, such as Action Verbs (AV), Linking Verbs (LV) and Auxiliary Verbs (AuxV).

In addition to the English grammar terms, domain-specific terms (i.e. related to question-answering) where identified, which correspond to topics – these are listed in Table 3.

In Table 4 the three different levels of detail related to the Grammatical categories are presented to enable us to establish the influence of the different levels of detail on the classification performance; a list of all the Grammatical categories and corresponding acronyms is displayed in the Appendix.

Table 3: Domain Specific Terms Categories

Category Name	**Terms Example**
Health	Specific Terms related to health, medicine, beauty.
Sports	Game and recreation, sports events, sports.
Arts and entertainment	Entertainment , Celebrities Name, lyrics, Movies, Books, Authors
Food and drinks	Foods, Drinks, Recipes
Animals	Pets, wild animals.
Science and math	Specific Terms related to Science, math.
Technology and internet	Software and Applications, Site, Website, URL, Database and Servers.
Society and culture	Environment, Holidays, Months, history, political, Relationships, Family.
News and events	Newspapers, Magazines, Documents, events
Job, Education and Reference	Careers, Institutions, Associations, Clubs, Parties, Foundations and Organizations.
Business and Finance	Money, company,products, Economy.
Travel and places	Geographical Areas, Transportation, Places and Buildings, Countries.

Table 4: The three levels taxonomy

Levels	Description	Classes
S	Consists of All Phrase classes	$NP, VP, PP, AP, AdvP.$
Level L1	Consists of the seven main word classes and Question words	$N, V, Adj, Adv, Conj, D, P, QW$
Level L2	Consists of the word classes sub classes	$CN, PN, NN, Pron, AV, LV, AuxV$
Level L3	Consists of all the specific classes that were created for the question classification	NN_C, NN_O, QW_{Who}, QW_{What}, QW_{Where}, QW_{When}, QW_{How}, QW_{Which}, PN_C, PN_S, PN_{HLT}, PN_{HMD}, PN_R, PN_{HN}, PN_{SA},PN_{BN}, PN_E, PN_{Ent}, PN_{BDN}, PN_G, PN_{IOG}, PN_{PB}, PN_{CO}, CN_A, CN_{SWU}, CN_{HN}, CN_{OS}, CN_{OP}, CN_{HLT}.

4.1.4. Constructing Term Taxonomy

In order to construct term categories the following steps have been taken using a Java program that has been developed by [35] and updated by [33] and [34]:

1. Parse and automatically extract terms from each question.
2. Automatically map terms to their POS tag, e.g. *"what is the capital of Spain"* is mapped as: *"What − > QW"*, *"is − > LV"*, *"the − > D"*, *"capital − > N"*, *"of − > P"* and *"Spain −− > N"*.
   after tagging each term to one of the main word classes mentioned above, a further tagging is done to assign each term to its sub-class if applicable. For example, *"What"*, *"is"* and *"the"* will not be mapped to any further categories, *"capital"* will be mapped to *"CN"*, *"of"* will not be mapped to any further categories and *"Spain"* will be mapped to *"PN"*.
3. Finally, after each term is mapped to one of the word classes, it will be mapped to the domain specific term category; the proposed categories were created after the analysis of the selected datasets. A detailed explanation of each category is provided in the appendix. For example, *"What"* will be mapped to − > QW_{what}", *"is"* and *"the"* will not be mapped to any further categories, will not be mapped to any further categories, *"capital"* will be mapped to "CN_{OS}", *"of"* will not be mapped to any further categories and *"Spain"* will be mapped to "PN_G".

The final step has resulted in the final refined taxonomy of term categories . Our tag-set contains all terms extracted from the dataset that we have used. We also added all possible terms in all the 7 main word classes except the Proper

Noun Category, since Proper Nouns are infinite. Note that although our solution does not require knowing all Proper Nouns, it is still capable of classifying text that contain unrecognized Proper Nouns.

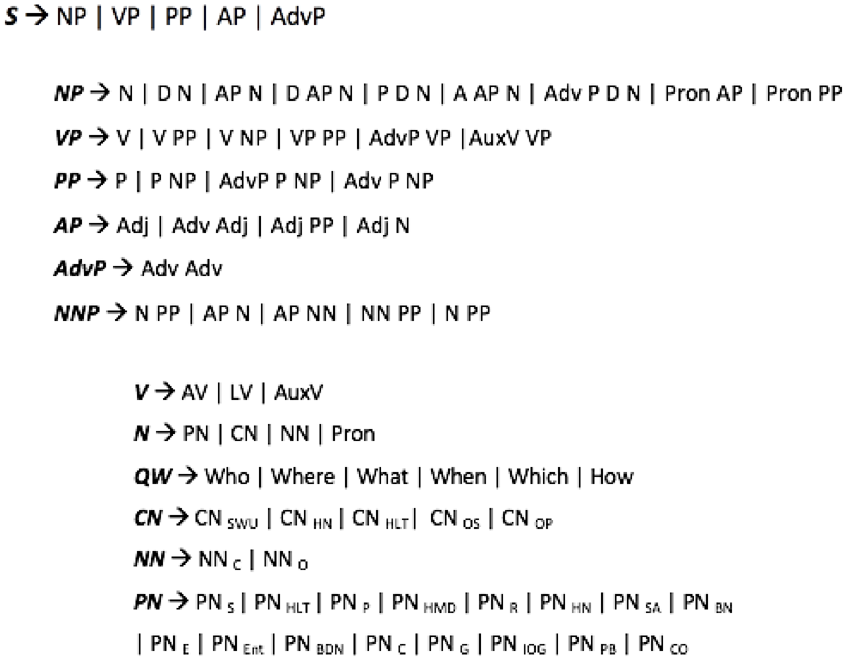
4.1.5. Grammar

In this section we present the formal grammar rules adopted in this study which are used for term mapping; these rules are based on the Context-Free Grammar (CFG).

A grammar is a tuple (N, Σ, P, S), where:

1. N is a finite set of non-terminal symbols, which can be single words, such as *"poems"*, or groups of words such as *"Emily Bront"* or *"United Nations"*;
2. Σ is a finite set of terminal symbols that is disjoint from N (i.e Σ and N have no common elements); the terminal symbols are the grammatical categories both general and domain-specific (e.g. noun, verb, proper noun, action verb);
3. P is a finite set of production rules of the form $(\Sigma \cup N)^* N (\Sigma \cup N)^* \rightarrow (\Sigma \cup N)^*$, and
4. $S \in N$ is the starting symbol.

Below we illustrate in Fig. 2 a number of rules which show how the grammatical categories are derived, starting from the highest level (the starting symbol, i.e. the question) to the lowest level of detail (level 3). The grammar rules contain grammatical features such as phrases, verbs, nouns, plurals and questions.



S → NP | VP | PP | AP | AdvP

NP → N | D N | AP N | D AP N | P D N | A AP N | Adv P D N | Pron AP | Pron PP
VP → V | V PP | V NP | VP PP | AdvP VP |AuxV VP
PP → P | P NP | AdvP P NP | Adv P NP
AP → Adj | Adv Adj | Adj PP | Adj N
AdvP → Adv Adv
NNP → N PP | AP N | AP NN | NN PP | N PP

V → AV | LV | AuxV
N → PN | CN | NN | Pron
QW → Who | Where | What | When | Which | How
CN → CN_{SWU} | CN_{HN} | CN_{HLT}| CN_{OS} | CN_{OP}
NN → NN_{C} | NN_{O}
PN → PN_{S} | PN_{HLT} | PN_{P} | PN_{HMD} | PN_{R} | PN_{HN} | PN_{SA} | PN_{BN}
| PN_{E} | PN_{Ent} | PN_{BDN} | PN_{C} | PN_{G} | PN_{IOG} | PN_{PB} | PN_{CO}

Figure 2: Grammar Rules

4.2. Phase II: Parsing and Mapping

In Phase II, the question is transformed into a pattern of grammatical terms by first parsing the question and then mapping each term to its grammar terminals, as illustrated in Algorithm 1. For sentence such as *"list of movies"* the parsing and mapping is simple since it contains only single words; each word is parsed and mapped individually and will be transformed into the following pattern $[CN_{OS} + P + CN_{OP}]$.

However, for question such as *"What did Alessandro Volta invent in 1800?"* which contains both single and compound words, first compound words will be parsed and extracted then single words, terms will be extracted as follow; *”What”*, *”did”*, *”Alessandro Volta”*, *”invent”*, *”in”*, *”1800”* and the question will be transformed into the following pattern $[QW_{What} + Auxv + PN_C + AV + P + NN_C]$. Some questions or sentences might contain compound words which consist of more than three terms, For example, in a sentence like *”University of Portsmouth Library”* terms will be extracted as follow; *”University of Portsmouth”* will be parsed as one word since it is a compound word and *”Library”* will be parsed as a single word. The following pattern will be formulated $[PN_{IOG} + CN_{OS}]$.

Another example is illustrated in Fig. 3 for the question *’what are the symptoms of diabetes?’*. The right-hand side of the figure illustrates the parsing of the question to extract the set of terms using the proposed grammatical rules discussed in Section 4.1.5, while the left-hand side illustrates the mapping of the terms to the grammar non-terminals. As a result of this process, the example question is transformed into the following pattern: $[QW_{What} + LV + D + CN_{OP} + P + CN_{HLT}]$. In this given example the pattern is a representation of the grammatical pattern in level 3 (i.e. the most detailed level).

Algorithm 1 Parsing and Mapping Algorithm

```
Read question $Q$ from input file. {$Q$ is the set of all questions in the dataset.}
Read grammar rules and store it in $G$ {$G$ is the set of grammar rules.}
for each $q_i$ in $Q$ do
    Parse $q_i$ and extract the set of terms $t_i$ {$t_i$ is a set that contains all the words in $q_i$ }
    for each $w_j$ in $t_i$ do
        $c_k$ = Map($w_j$, $G$) {This maps term $w_j$ based on $G$ into category $c_i$}
        if $c_k$ is null then
            $c_k = PN$ {If no category found for term $w_j$, assume it is a proper noun.}
            if $c_{k-1}$ is $PN$ then
                $combine(c_{k-1}, c_k)$ {Replace any number of consecutive $PN$ with a single $PN$}
            end if
        end if
    end for
end for
```

4.3. Phase III: Question Classification

In this phase the patterns generated in Phase II are used for machine learning, the aim of this phase is to build a model for automatic classification. The classification is done by following the standard process for machine learning, which involves the splitting of the dataset into a training dataset and a test dataset.

The training dataset is used for building the model, and the test dataset is used to evaluate the performance of the model. Once a model of satisfactory performance has been identified, it can be used for the classification on unlabelled questions.

5. Experimental Study and Results

The objective of the experimental study is to investigate the ability of machine learning classifiers to distinguish between different question types based on the different levels of detail used in the term taxonomy.

Four machine learning algorithms, were used for question classification. These are briefly described below.

1. The Decision Tree (DT) is a method for approximating discrete-valued functions that is robust to noisy data and capable of learning disjunctive expressions. It classifies instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. [32]. J48 and RandomForest are two of the widely used Decision Tree algorithms.
    (a) J48 Decision tree (J48) is an extension of the ID3 algorithm and is typically used in the machine learning and natural language processing domains [43]. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges and derivation of rules. In the WEKA data mining tool, J48 is an implementation of the C4.5 algorithm [44].

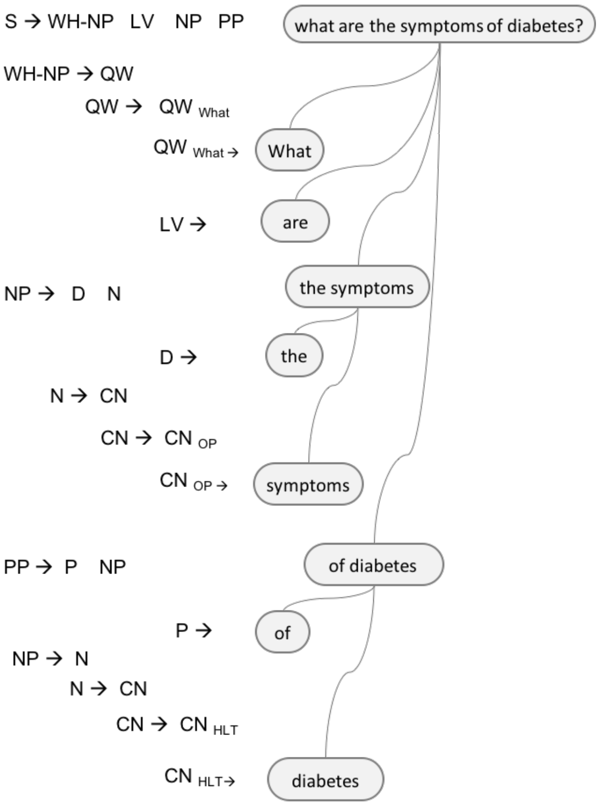


Figure 3: Phase II: Parsing and Mapping Example

(b) Random forests (RF) are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [3], [16].

2. Naive Bayes (NB) estimates the parameters of a multinomial generative model for instances, then finds the most probable class for a given instance using the Bayes rule and the Nave Bayes assumption that the features occur independently of each other inside a class [45]. In practice the Nave Bayes learner performs remarkably well in many text classification problems [32] and is often used as a baseline in text classification because it is fast and easy to implement. Less erroneous algorithms tend to be slower and more complex [45].
3. In Support Vector Machine (SVM) input vectors are non-linearly mapped to a very high-dimension feature space. In this feature space a linear decision surface is constructed. Special properties of the decision surface ensures the high generalization ability of the learning machine [7]. SVMs are helpful in text categorization as their application can significantly reduce the need for labeled training instances in both the standard inductive and transductive settings. In addition, SVMs have the ability to generalize well in high-dimensional feature spaces. SVMs eliminate the need for feature selection making the application of text categorization considerably easier and do not require any parameter tuning since they can find good parameter settings automatically [18].

We used 1,160 questions that were randomly selected from the datasets that were mentioned in Section 4.1.2: Yahoo Non-Factoid Question, TREC 2007 Question Answering Data and a Wikipedia dataset. Their distribution is given in Table 5.

Table 5: Data distribution

Question type	Total
Causal	31
Choice	12
Confirmation	321
Factoid	688
Hypothetical	7
List	101

To assess the performance of the machine learning classifiers, the Weka[4] software [10] was used. The experiments were set up using the typical 10-fold cross validation, i.e. the dataset is split into 10 folds, and each fold in used, in turn, for testing, while the other 9 are used for training. The output of the training process is a model, which is then used to classify the questions in the test fold. The labels produced by the model are matched to the true labels and typical performance indicators, such as accuracy, precision, recall, and F-score, are calculated. The results are presented in the next subsection.

5.1. Results

In this section we present and analyse the results of the machine learning algorithms for each of the three levels of the term taxonomy.

5.1.1. Level-1

Table 6 presents classification performance details (Precision, Recall and F-Measure) of the J48, Naive Bayes, RandomForest and SVM classifiers. Results show that Decision Tree (J48) identified correctly (i.e. Recall) 86.6% of the questions, while SVM identified correctly 85.3% of the questions, RandomForest 84.7% and Naive Bayes 81.6%.

More specifically, looking at where the errors occur, J48 could not identify causal question and misclassified 3.2% as Confirmation and 96.8% were misclassified as Factoid. For the choice questions J48 misclassified 41.7% as Confirmation and 25% as Factoid. From the Confirmation questions 1.6% were misclassified as hypothetical. Furthermore, 0.7% of the Factoid questions were misclassified as Confirmation, 0.15% as Causal, 0.15% as Choice and 1.2% as List. For the List Type of question, 1% were of the questions were misclassified as Confirmation and 87.1% were misclassified as Factoid. Moreover, J48 could not identify Hypothetical questions and incorrectly classify them as Factoid.

Naive Bayes classifier incorrectly classified 3.2% of the causal questions as Choice, 87.1% as Factoid, 3.2% as Confirmation and 6.5% as Hypothetical. In addition, NB could not identify Choice questions and misclassified 42% as Confirmation, 42% as Factoid and 16% as List. Furthermore, 1.2% of the Confirmation questions were misclassified as Choice, 3.4% as Factoid, 2.1% as Hypothetical and 0.3% as List. For the Factoid questions, 1.1% were misclassified as Causal, 2% as Choice, 1.7% as Confirmation, 2% as Hypothetical and 3% as List. Moreover, 14% of the Hypothetical questions were misclassified as Causal and 57% as Factoid. For the List Type of question NB incorrectly classified 3% as Confirmation and 86% as Factoid.

Similar to NB classifier, RandomForest Classifier could not identify causal and choice questions. For the causal NB incorrectly classified 3.2% as Confirmation and 96.8% as Factoid. Moreover, 41.7% of the Choice questions were misclassified as Confirmation and 58.3% as Factoid. For the Confirmation questions, 0.3% were misclassified as Choice and 2.8% as Hypothetical. Moreover, 0.6% of the Factoid questions were misclassified as Causal, 0.3% as Choice, 0.9% as Confirmation and 3.2% as List. RandomForest Could not identify Hypothetical questions misclassified them as Factoid. In addition, 2% of the List questions were misclassified as Confirmation and 80% as Factoid. Finally, the Support Vector Machine classifier could not identify Causal questions and 3.2% of the questions were misclassified as Confirmation and 96.8% were misclassified as Factoid. From the choice questions 33.3% were misclassified as confirmation and 33.3% were misclassified as factoid. Similarly, 1% of the list questions were misclassified as confirmation and 92% were misclassified as factoid. Moreover, 2.8% of Confirmation questions

[4] http://www.cs.waikato.ac.nz/ml/weka/

Table 6: Performance of the classifiers in Level (1) - Best results are highlighted in bold, the * indicates that the results are significantly better.

	J48			*SVM*			*RF*			*NB*		
Accuracy:	**86.6%***			85.3 %			84.7%			81.6%		
Precision:	**0.822**			0.8			0.796			0.779		
Recall:	**0.866**			0.853			0.847			0.816		
F-score	**0.826**			0.813			0.814			0.79		
Class:	P	R	F	P	R	F	P	R	F	P	R	F
Causal	0	0	0	0	0	0	0	0	0	0	0	0
Choice	**0.8**	**0.333**	**0.471**	0.5	**0.333**	0.40	0	0	0	0	0	0
Conf.	**0.963**	**0.984**	0.974	**0.963**	0.966	0.964	0.957	0.969	0.963	0.934	0.928	0.931
Factoid	**0.835**	**0.978**	**0.901**	0.826	0.967	0.891	0.83	0.951	0.886	0.826	0.924	0.872
Hypo.	0	0	0	**1**	**0.429**	**0.6**	0	0	0	0.083	0.286	0.129
List	**0.6**	0.119	0.198	0.368	0.069	0.117	0.45	**0.178**	**0.255**	0.344	0.109	0.165

were misclassified as factoid and less than 1% were misclassified as choice. For the factoid questions 0.4% were misclassified as Causal, 0.3% were misclassified as Choice, 0.9% were misclassified as confirmation and 1.7% were misclassified as List. In addition, most of the hypothetical questions, i.e. 57%, were misclassified as factoid.

5.1.2. Level-2

Table 7 presents classification performance details (Precision, Recall and F-Measure) of the J48, Naive Bayes, RandomForest and SVM classifiers for level 2. Results show that Decision Tree (J48) identified correctly (i.e. Recall) 87.2% of the questions, while SVM identified correctly 86.6% of the questions, RandomForest 85.8% and Naive Bayes 81.9%.

More specifically, in this level J48 could not identify causal and Choice questions and misclassified 3.2% for the Causal questions as Confirmation and 96.8% were misclassified as Factoid. For the choice questions J48 misclassified 42% as Confirmation and 58% as Factoid. From the Confirmation questions 1.6% were misclassified as hypothetical and 0.6% as List. Furthermore, 0.9% of the Factoid questions were misclassified as Confirmation, 0.15% as Causal and also 0.15% as List. For the List Type of question, 1% were of the questions were misclassified as Confirmation and 81% were misclassified as Factoid. Moreover, J48 could not identify Hypothetical questions and incorrectly classified them as Factoid.

Similar to J48 classifier, Naive Bayes classifier could not identify Causal and Choice questions and incorrectly classified 3.2% of the causal questions as Confirmation, 90.3% as Factoid, and 6.5% as Hypothetical. For the Choice questions 42% were misclassified as Confirmation and 58% as Factoid. Furthermore, 1.5% of the Confirmation questions were misclassified as Choice, 2.5% as Factoid, 3.1% as Hypothetical and 1.9% as List. For the Factoid questions, 1.9% were misclassified as Causal, 0.3% as Choice, 1.3% as Confirmation, 1.6% as Hypothetical and 2% as List. Moreover, 14% of the Hypothetical questions were misclassified as Causal and 57% as Factoid. For the List Type of question NB incorrectly classified 5% as Confirmation and 77% as Factoid.

RandomForest Classifier incorrectly classified 3.5% of the causal questions as Confirmation and 90.3% as Factoid. Moreover, similar to NB and J48 classifiers, RF could not identify Choice questions and misclassified 42% as Confirmation and 58% as Factoid. For the Confirmation questions, 0.6% were misclassified as Choice and 2.5% as Factoid. Moreover, 0.2% of the Factoid questions were misclassified as Choice, 1.2% as Confirmation and 1.6% as List. For the Hypothetical questions RF misclassified most of them 85.7% as Factoid. In addition, 3% of the List questions were misclassified as Confirmation and 83% as Factoid.

Finally, using Support Vector Machine, 3.2% of the Causal questions were misclassified as Confirmation and 90.3% were misclassified as Factoid. SVM is the only classifier in this level to classify choice questions but misclassified 42% as confirmation and also 42% as factoid. Moreover, 1.2% of Confirmation questions were misclassified as factoid and less than 1% were misclassified as choice and list. For the factoid questions 1.3% were misclassified as Causal, 0.15% were misclassified as Choice and 0.15% as Hypothetical, 0.9% were misclassified as confirmation and 1.9% were misclassified as List. In addition, 14% of the hypothetical questions were misclassified as Causal and 43% as Factoid. Finally, 2% of the list questions were misclassified as confirmation and 71% were misclassified as

Table 7: Performance of the classifiers in Level (2) - Best results are highlighted in bold, the * indicates that the results are significantly better.

	J48			*SVM*			*RF*			*NB*		
Accuracy:	**87.2%** *			86.6 %			85.8%			81.9%		
Precision:	0.838			**0.841**			0.839			0.797		
Recall:	**0.872**			0.866			0.859			0.82		
F-score	0.832			**0.844**			0.821			0.801		
Class:	P	R	F	P	R	F	P	R	F	P	R	F
Causal	0	0	0	0.167	**0.065**	0.093	**1**	**0.065**	**0.121**	0	0	0
Choice	0	0	0	**0.4**	**0.167**	**0.235**	0	0	0	0	0	0
Conf.	**0.96**	0.978	**0.969**	0.957	0.975	0.966	0.948	0.969	0.958	0.936	**0.91**	0.923
Factoid	0.838	**0.988**	**0.907**	**0.855**	0.956	0.903	0.834	0.971	0.897	0.836	0.929	0.88
Hypo.	0	0	0	0.75	**0.429**	**0.545**	**1**	0.143	0.25	0.08	0.286	0.125
List	**0.857**	0.178	0.295	0.643	**0.267**	**0.378**	0.56	0.139	0.222	0.474	0.178	0.259

Factoid.

5.1.3. Level-3

Table 8 presents the classification performance details (Precision, Recall and F-Measure) of the J48, Naive Bayes, RandomForest and SVM classifiers for level 3. Results show that Decision Tree (J48) identified correctly (i.e. Recall) 90.1% of the questions, while SVM identified correctly 88.6% of the questions, Naive Bayes 83.5% and RandomForest 87.7%.

More specifically, looking at where the errors occur, when using J48 3.2% of the causal questions were misclassified as Confirmation and 12.9% were misclassified as Factoid. For the choice questions J48 could not identify this type of question and misclassified 41.1% as Confirmation and 58.3% as Factoid. From the Confirmation questions 0.31% were misclassified as Causal, 0.62% as Factoid and also 0.62% as List. Furthermore, 0.7% of the Factoid questions were misclassified as Confirmation, 1% as Causal, 1% as List and 0.4% as Hypothetical.

For the List Type of question, 1% of the questions were misclassified as Confirmation and 60.4% were misclassified as Factoid. Moreover, J48 could not identify Hypothetical questions and incorrectly classify them as Factoid.

The Naive Bayes classifier incorrectly classified 6.5% of the causal questions as Confirmation, 80.6% as Factoid and 3.2% as List. Similar to J48 classifier, NB could not identify Choice questions and misclassified 41.7% as Confirmation and 58.3% as Factoid. Furthermore, 0.9% of the Confirmation questions were misclassified as Choice, 3.4% as Factoid, 2% as Hypothetical and 0.9% as List.

For the Factoid questions, 1.3% were misclassified as Causal, 0.43% as Choice, 2.5% as Confirmation, 0.87% as Hypothetical and 2.2% as List. Moreover, 14.3% of the Hypothetical questions were misclassified as Causal and 57.1% as Factoid. For the List Type of question NB incorrectly classified 7% as Confirmation and 65.3% as Factoid. RandomForest classifier incorrectly classified 6.4% of the causal questions as Confirmation and 58.3% as Factoid. Similar to J48 and NB classifiers, RF could not identify Choice questions and misclassified 41.7% as Confirmation and 58.3% as Factoid. For the Confirmation questions, 0.6% were misclassified as Choice and 3.4% as Factoid. Moreover, 1.2% of the Factoid questions were misclassified as Confirmation and 0.7% as List. Hypothetical questions were 71.4% misclassified as Factoid. In addition, 2% of the List questions were misclassified as Confirmation and 72.3% as Factoid.

Finally, using Support Vector Machine, 3.2% of the causal questions were misclassified as Confirmation and 32.2% were misclassified as Factoid. From the choice questions 41.7% were misclassified as confirmation and 33.3% were misclassified as factoid. Similarly, 4% of the list questions were misclassified as confirmation and 45.5% were misclassified as factoid. These results indicate that SVM could not distinguish between causal, choice and list types of questions and incorrectly classified most of them as confirmation and factoid questions. Moreover, 1.6% of Confirmation questions were misclassified as factoid and less than 1% were misclassified as choice or list. For the factoid questions 4.6% were misclassified as list, 1.2% were misclassified as causal, 1% were misclassified as confirmation and less than 1% were misclassified as choice. In addition, most of the hypothetical questions 57.1% were misclassified as factoid.

Table 8: Performance of the classifiers in Level (3) - Best results are highlighted in bold, the * indicates that the results are significantly better.

	J48			*SVM*			*RF*			*NB*		
Accuracy:	**90.1 %***			88.6%			87.7%			83.5%		
Precision:	**0.885**			0.88			0.872			0.814		
Recall:	**0.901**			0.886			0.877			0.835		
F-score:	**0.885**			0.881			0.85			0.818		
Class:	P	R	F	P	R	F	P	R	F	P	R	F
Causal	0.722	**0.839**	**0.776**	0.714	0.645	0.678	**1**	0.194	0.324	0.231	0.097	0.136
Choice	0	0	0	**0.429**	**0.25**	**0.316**	0	0	0	0	0	0
Conf.	**0.963**	**0.984**	**0.974**	0.948	0.972	0.96	0.948	0.96	0.954	0.906	0.928	0.917
Factoid	0.891	0.965	**0.927**	**0.903**	0.929	0.915	0.85	**0.981**	0.911	0.85	0.927	0.887
Hypo.	0	0	0	**1**	**0.429**	**0.6**	**1**	0.286	0.444	0.133	0.286	0.182
List	0.813	0.386	0.523	0.6	**0.505**	**0.548**	**0.839**	0.257	0.394	0.609	0.277	0.381

6. Discussion

The results show (Figure 4) that with each level there is an improvement in the results when moving from *level 1* to *level 2* and from *level 2* to *level 3*. The improvement in the performance from *level 1* to *level 2* is marginal and there is an increase in the performance from *level 2* to *level 3*. In addition, the Results indicate that J48 significantly preformed better than SVM, RF and NB in all three levels. We used Weka corrected t-test with the threshold value of 0.05 (i.e. p-value <0.05).

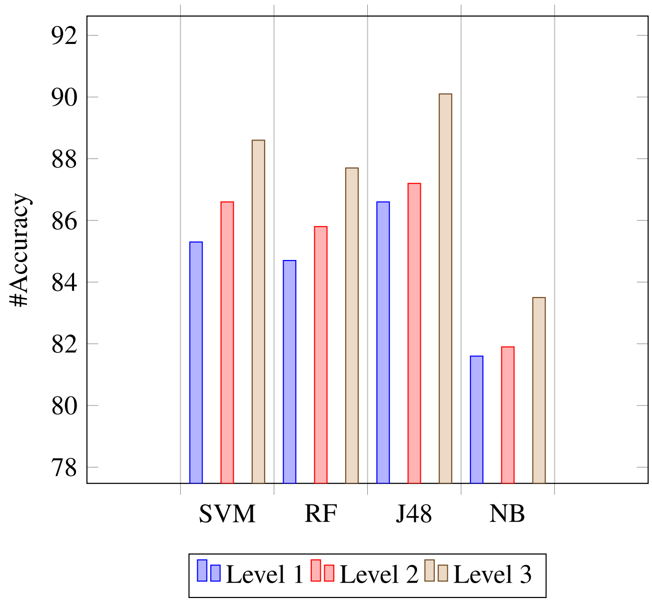


Figure 4: Accuracy of the classifiers in level 1, 2 and 3

Level 1 and level 2 contain general grammatical categories of the English language. When only the higher level categories are used (i.e. level 1), while there are variations between the different learning algorithms, the overall picture is that the best performance occurs for confirmation and factoid questions and the worst performance for Causal questions. In this level, all of the classifiers could not identify at least one question type. SVM could not identify causal questions, RF could not identify causal, choice and hypothetical questions, J48 could not identify causal and hypothetical questions and NB could not identify causal and choice questions.

When subcategories of the English main syntactic categories are used, i.e. level 2, we see a dramatic improvement in the performance of all classifiers. SVM is the only classifier that could identify all type of questions but similar to the performance in level 1 NB, J48 and RF could not identify at least one question type. RF could not identify Choice questions, but for the causal and hypothetical questions RF has a recall of 1 for these classes which indicates that there are no false positives, i.e. all instances identifies by the models as causal or hypothetical are truly these two types. Furthermore, J48 could not identify choice and hypothetical questions in addition to causal, in contrast to level 1 in which J48 was able to classify choice questions. NB also could not identify causal and choice questions.

The sub-categories at level 2 have also marginally improved the performance for the some question types like list for the classifiers SVM, J48 and NB, factoid for RF and J48, hypothetical for RF, Causal for the classifiers SVM and RF and finally confirmation for the SVM classifier. Level 3, which includes the domain-specific grammatical categories, led to significant improvements of the performance of all classifiers. In this level J48 and NB could identify causal questions. Regarding Choice question, RF, J48 and NB still could not identify this type of question; in addition, similar to level 1 and 2, J48 with the more detailed grammar could not identify hypothetical questions. SVM is the only classifier that could identify and classify all type of questions. These results indicate that the 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.

Although, we were unable to perform direct comparisons with other state-of-the-art methods – this is because different methods use different datasets with different set labels and the implementation and testing details of these methods are not available – in this section we discuss the performance of state-of-the-art methods in terms of accuracy.

[25] proposed a hierarchical classifier that classifies questions into fine grained classes, using Sparse Network of Winnows (SNoW); the proposed approach achieved accuracy of 92.5% for coarse grained classes and 85% for fine grained classes when using only syntactical features; after adding semantic features the accuracy accounted for 89.3%.

Most previous works were based on Li and Roth classification of question and deals with factoid questions only. [54] used bag-of-words features on different machine learning algorithms; SVM performed better comparing with the other classifiers like kNN and NB. SVM achieved an accuracy of 80.2% with fine grained classes and 85.8% with coarse grained classes. [17] used head word features and wordNet in addition to unigrams; their liner SVM and maximum entropy models reach the accuracy of 89.2% and 89% respectively. The statistical classifier in [30] is based on SVM and achieved an accuracy of 90.2% using coarse grained classes and 83.6% using fine grained classes. [23] classified factoid questions using head Noun tagging combining syntactical and semantic features; they uses conditional random fields (CRFs) and SVM; the model achieved an accuracy of 85.6%. [39] proposed an approach which is based on question patterns and designed features; they achieved an accuracy of 95.2% and 91.6% for coarse grained classes and fine grained classes respectively using SVM.

Even-though these approaches achieved good accuracy rate, they have used Li and Roth classification of questions, which is based on a large number of categories. In addition, this classification only focuses on solving the problem of classifying and identifying factoid types of question. Furthermore, the majority of these previous works used SVM for the classification process; in our experiments it has been shown that other classifiers like J48 could be used for the classification with good results.

Furthermore, [6] classified open ended questions using SVM and achieved an accuracy of 74.6% on average. However, the data in this work were collected from textbook and references, which are not representative of questions typically asked in question-answering systems. In addition, some of the data attributes have been removed like stop words,"s" for plural words and "ly" for adverbs, which we believe are important to identify question types. For example, plural terms are one of the main identification feature of question type "List".

In conclusion, our approach outperforms the previous ones due to the ability of our approach to classify different questions types, not just factoid. In addition, our approach uses domain-specific information which facilitate the identification of domain categories, unlike previous works which focus only on the type of question.

7. Conclusion and Future Work

In this paper we proposed a Grammar Based Framework for question categorization and classification (GQCC) to automatically classify questions through machine learning by taking advantage of domain-specific information and by preserving the structure of the questions. For the later purpose, a new representation was proposed, in which the question is represented as a grammatical pattern, i.e. a pattern formed of grammatical categories corresponding to the terms in the text. To transform the text into this representation we proposed a formal grammar-based approach. The results show that our solution led to a good performance in classifying questions.

Despite that the proposed framework was tested on question classification, it can be applied to other text classification problems, which will be investigated as future work. In addition, we aim at examining and analyzing more questions from different data-sets and extending the analysis of the different types of question. Furthermore, we will test our approach using different data-sets with existing labels e.g (Li and Roth) question classification [25]. Also, most questions dataset suffers form class imbalance between the labels, we aim to investigate the use of ensemble learning and other data mining methods to deal with the class imbalance problem [41, 1].

References

[1] Bader-El-Den, M., Teitei, E., Adda, M., 2016. Hierarchical classification for dealing with the class imbalance problem, in: Neural Networks (IJCNN), 2016 International Joint Conference on, IEEE. pp. 3584–3591.

[2] Benamara, F., 2004. Cooperative question answering in restricted domains: the webcoop experiment.

[3] Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

[4] Broder, A., 2002. A taxonomy of web search. ACM Sigir forum 36, 3–10.

[5] Bu, F., Zhu, X., Hao, Y., Zhu, X., 2010. Function-based question classification for general qa, in: Proceedings of the 2010 conference on empirical methods in natural language processing, Association for Computational Linguistics. pp. 1119–1128.

[6] Bullington, J., Endres, I., Rahman, M., 2007. Open ended question classification using support vector machines. MAICS 2007 .

[7] Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20, 273–297.

[8] Greenbaum, S., Nelson, G., 2002. An introduction to English grammar. Pearson Education.

[9] Hacioglu, K., Ward, W., 2003. Question classification with support vector machines and error correcting codes, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2, Association for Computational Linguistics. pp. 28–30.

[10] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The weka data mining software: an update. ACM SIGKDD explorations newsletter 11, 10–18.

[11] Hao, T., Xie, W., Wu, Q., Weng, H., Qu, Y., 2017. Leveraging question target word features through semantic relation expansion for answer type classification. Knowledge-Based Systems 133, 43–52.

[12] Hao, T., Xie, W., Xu, F., 2015. A wordnet expansion-based approach for question targets identification and classification, in: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer, pp. 333–344.

[13] Hardy, H., Cheah, Y.N., 2013. Question classification using extreme learning machine on semantic features. Journal of ICT Research and Applications 7, 36–58.

[14] Hasan, A.M., Zakaria, L.Q., 2016. Question classification using support vector machine and pattern matching. Journal of Theoretical and Applied Information Technology 87, 259.

[15] Higashinaka, R., Isozaki, H., 2008. Corpus-based question answering for why-questions.

[16] Ho, T.K., 1995. Random decision forests, in: Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, IEEE. pp. 278–282.

[17] Huang, Z., Thint, M., Qin, Z., 2008. Question classification using head words and their hypernyms, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. pp. 927–936.

[18] Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. Machine learning: ECML-98 , 137–142.

[19] King, M., 1983. Parsing natural language. Academic Press London.

[20] Kolomiyets, O., Moens, M.F., 2011. A survey on question answering technology from an information retrieval perspective. Information Sciences 181, 5412–5434.

[21] Le-Hong, P., Phan, X.H., Nguyen, T.D., 2015. Using dependency analysis to improve question classification, in: Knowledge and Systems Engineering. Springer, pp. 653–665.

[22] Leech, G., Svartvik, J., 2013. A communicative grammar of English. Routledge.

[23] Li, F., Zhang, X., Yuan, J., Zhu, X., 2008. Classifying what-type questions by head noun tagging, in: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics. pp. 481–488.

[24] Li, X., Huang, X.J., WU, L.d., 2005. Question classification using multiple classifiers, in: Proceedings of the 5th Workshop on Asian Language Resources and First Symposium on Asian Language Resources Network.

[25] Li, X., Roth, D., 2006. Learning question classifiers: the role of semantic information. Natural Language Engineering 12, 229–249.

[26] Li, Y., Su, L., Chen, J., Yuan, L., 2017. Semi-supervised learning for question classification in cqa. Natural Computing 16, 567–577.

[27] Liu, Z., Jansen, B.J., 2017. Identifying and predicting the desire to help in social question and answering. Information Processing & Management 53, 490–504.

[28] Liu, Z., Jansen, B.J., 2018. Questioner or question: Predicting the response rate in social question and answering on sina weibo. Information Processing & Management 54, 159–174.

[29] May, R., Steinberg, A., 2004. Al, building a question classifier for a trec-style question answering system. AL: The Stanford Natural Language Processing Group, Final Projects .

[30] Metzler, D., Croft, W.B., 2005. Analysis of statistical question classification for fact-based questions. Information Retrieval 8, 481–504.

[31] Mishra, M., Mishra, V.K., Sharma, H., 2013. Question classification using semantic, syntactic and lexical features. International Journal of Web & Semantic Technology 4, 39.

[32] Mitchell, T.M., 1997. Machine learning. McGraw hill.

[33] Mohasseb, A., Bader-El-Den, M., Kanavos, A., Cocea, M., 2017a. Web queries classification based on the syntactical patterns of search types, in: International Conference on Speech and Computer, Springer. pp. 809–819.

[34] Mohasseb, A., Bader-El-Den, M., Liu, H., Cocea, M., 2017b. Domain specific syntax based approach for text classification in machine learning context, in: 2017 International Conference on Machine Learning and Cybernetics (ICMLC), IEEE Systems, Man and Cybernetics. pp. 658–663.

[35] Mohasseb, A., El-Sayed, M., Mahar, K., 2014. Automated identification of web queries using search type patterns., in: WEBIST (2), pp. 295–304.

[36] Mohd, M., Hashmy, R., 2018. Question classification using a knowledge-based semantic kernel, in: Soft Computing: Theories and Applications. Springer, pp. 599–606.

[37] Moldovan, D., Paşca, M., Harabagiu, S., Surdeanu, M., 2003. Performance issues and error analysis in an open-domain question answering system. ACM Transactions on Information Systems (TOIS) 21, 133–154.

[38] Nguyen, T.T., Nguyen, L.M., Shimazu, A., 2007. Improving the accuracy of question classification with machine learning, in: Research, Innovation and Vision for the Future, 2007 IEEE International Conference on, IEEE. pp. 234–241.

[39] Nguyen, T.T., Nguyen, L.M., Shimazu, A., 2008. Using semi-supervised learning for question classification, Information and Media Technologies Editorial Board. pp. 112–130.

[40] Nijholt, A., 1980. Context-free grammars: covers, normal forms, and parsing. 93, Springer Science & Business Media.

[41] Perry, T., Bader-El-Den, M., Cooper, S., 2015. Imbalanced classification using genetically optimized cost sensitive classifiers, in: Evolutionary Computation (CEC), 2015 IEEE Congress on, IEEE. pp. 680–687.

[42] Peters, R.A., 1968. A linguistic history of English. Houghton Mifflin.

[43] Quinlan, J.R., 1986. Induction of decision trees. Machine learning 1, 81–106.

[44] Quinlan, J.R., 2014. C4.5: programs for machine learning. Elsevier.

[45] Rennie, J.D., Shih, L., Teevan, J., Karger, D.R., et al., 2003. Tackling the poor assumptions of naive bayes text classifiers, in: ICML, Washington DC). pp. 616–623.

[46] Rose, D.E., Levinson, D., 2004. Understanding user goals in web search, in: Proceedings of the 13th international conference on World Wide Web, ACM. pp. 13–19.

[47] Sagara, T., Hagiwara, M., 2014. Natural language neural network and its application to question-answering system. Neurocomputing 142, 201–208.

[48] Smith, N.A., Heilman, M., Hwa, R., 2008. Question generation as a competitive undergraduate course project, in: Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge.

[49] Song, W., Wenyin, L., Gu, N., Quan, X., Hao, T., 2011. Automatic categorization of questions for user-interactive question answering. Information Processing & Management 47, 147–156.

[50] Ture, F., Jojic, O., 2016. Simple and effective question answering with recurrent neural networks. arXiv preprint arXiv:1606.05029 .

[51] Van-Tu, N., Anh-Cuong, L., 2016. Improving question classification by feature extraction and selection. Indian Journal of Science and Technology 9.

[52] Xu, S., Cheng, G., Kong, F., 2016. Research on question classification for automatic question answering, in: Asian Language Processing (IALP), 2016 International Conference on, IEEE. pp. 218–221.

[53] Yen, S.J., Wu, Y.C., Yang, J.C., Lee, Y.S., Lee, C.J., Liu, J.J., 2013. A support vector machine-based context-ranking model for question answering. Information Sciences 224, 77–87.

[54] Zhang, D., Lee, W.S., 2003. Question classification using support vector machines, in: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ACM. pp. 26–32.

Appendix:
Grammar terms and corresponding abbreviations

Category Name	Abbreviation
Verbs	V
Action Verbs	AV
Auxiliary Verb	$AuxV$
Linking Verbs	LV
Adjective	Adj
Adverb	Adv
Determiner	D
Conjunction	$Conj$
Preposition	P
Noun	N
Pronoun	$Pron$
Numeral Numbers	NN
Ordinal Numbers	NN_O
Cardinal Numbers	NN_C
Proper Nouns	PN
Celebrities Name	PN_C
Entertainment	PN_{Ent}
Newspapers, Magazines, Documents, Books	PN_{BDN}
Events	PN_E
Companies Name	PN_{CO}
Geographical Areas	PN_G
Places and Buildings	PN_{PB}
Institutions, Associations, Clubs, Foundations and Organizations	PN_{IOG}
Brand Names	PN_{BN}
Software and Applications	PN_{SA}
Products	PN_P
History and News	PN_{HN}
Religious Terms	PN_R
Holidays, Days, Months	PN_{HMD}
Health Terms	PN_{HLT}
Science Terms	PN_S
Common Noun	CN
Common Noun Other- Singular	CN_{OS}
Common Noun- Other- Plural	CN_{OP}
Database and Servers	CN_{DBS}
Advice	CN_A
Entertainment	CN_{Ent}
History and News	CN_{HN}
Site, Website, URL	CN_{SWU}
Health Terms	CN_{HLT}
Question Words	QW
How	QW_{How}
What	QW_{What}
When	QW_{When}
Where	QW_{Where}
Who	QW_{Who}
Which	QW_{Which}