Classification of Factoid Questions Intent Using Grammatical Features

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

In question-answering systems, question classification is a fundamental task. Identifying the accurate question type enhances the retrieval of more accurate answers, however, the continuing growth of the amount of web content makes the retrieval of relevant answers difficult. Factoid questions are the most challenging type of question to classify. Various approaches have been proposed with the objective of enhancing the identification and the classification of factoid questions; most of these are approaches based on semantic features and bag-of-words. In this paper, a Grammar-based framework for Questions Categorization and Classification (GQCC) is used. The framework makes use of three main features which are, grammatical features, domain specific features and patterns. These features utilize the structure of the questions. Using machine learning algorithms for the classification process, experimental results show that our approach has a good level of accuracy when comparing with the state-of-arts approaches.

Keywords:
Information Retrieval, Question Classification, Factoid Questions, Grammatical features, Machine Learning.

1. Introduction

In question and answering systems (QASs), questions classification is a fundamental task. Identifying the accurate question type enhances the retrieval of more accurate answers. However, the continuing growth of the amount of web content makes the retrieval of relevant answers difficult. Factoid questions are the most challenging type of question to classify. Various approaches have been proposed with the objective of enhancing the identification and the classification of factoid questions; most of these are approaches based on semantic features and bag-of-words. Several question taxonomies have been proposed [1, 2, 3, 4, 5]. The most popular classification taxonomy of factoid ('wh-') questions is Li and Roth's categories [5], which consists of a set of six coarse-grained categories and fifty fine-grained ones.

The classification of the questions performed in QASs directly affects the answers. Authors in [6] stated that most errors happen due to miss-classification of questions performed in QASs in which the task of generating answers to the users questions is directly related to the type of questions asked.

Classifying 'wh-' questions into proper semantic categories is found more challenging than classifying other types in question answering systems [7]. In addition, features are the key to obtain an accurate question classifier and linguistic features play an important role in developing an accurate question classifier [8]; recent studies classified users' questions using different features like bag-of-words [9], [10], [11] and uni-gram and word shape features [12]. Moreover, authors in [3] integrated pattern matching and machine learning techniques for the classification of questions, while [13] classified questions by their expected types of responses. According to [2] a question type is defined as a certain semantic category and is characterized by common properties.

Furthermore, machine learning algorithms have been used by many previous studies for the classification of questions. Support Vector Machine (SVM) is one of the most used algorithms [14], [4], [12], [15], [16],[17]. Combining different features such as syntactic, lexical and semantic attributes with a SVM classifier improves the classification accuracy [11]. Other works like [9] and [11] used other machine learning algorithms besides SVM such as Naive Bayes, Nearest Neighbors and Decision Tree.

In a previous study [1], a Grammar-based framework for Questions Categorization and Classification (GQCC) was proposed. In this study, the framework is applied to question classification according to Li and Roth's [5] categories of intent, using three main features which are: (1) Grammatical features (2) Domain specific features and (3) Patterns. These features transfer the question into a new representation which has the advantage of preserving the grammatical structure of the question. The aim is to assess the influence of using the structure of the question and the domain-specific syntactic categories and features on the classification performance. To achieve this aim, the following objectives are defined:

1. Investigate the influence of the different details of general grammatical features and domain-specific grammatical features on the classification performance;
2. Compare the performance of different machine learning algorithms for the classification of factoid questions intent;
3. Investigate the classification performance in comparison with state-of-the art approaches.

The rest of the paper is organised as follows. Section 2 out-

lines the categorization of questions and the previous work in question classification. Section 3 describes the proposed approach and the grammatical features used. The experimental setup and results are presented in Section 4, while the results are discussed in Section 5. Finally, Section 6 concludes the paper and outlines directions for future work.

2. Background

In this section we review previous work on question classification. Different proposed question categories are outlined in Section 2.1, while Section 2.2 reviews previous work on question classification methods.

2.1. Questions Categories

Different questions categories were proposed, which are summarised in Table 1. In [2] authors classified questions to eight types which are: list, definition, factoids, causal, relationship, hypothetical, procedural, and confirmation questions.

A list question requires as an answer a list of entities or facts. While, a definition question normally begins with "What is" and the objective is to find a definition of a term in the question. Furthermore, a factoid question usually starts with a Wh-questions such as What, When, Where and Who and the answer is usually a fact. A causal question starts with "Why" and requires explanation of an event.

In addition, in a relationship question the relation between two entities is ask, while in a procedural question a list of instructions for accomplishing the task mentioned in the question is required as an answer. The objective of a hypothetical question is to find information about a hypothetical event and has the form of "What would happen if" question. Finally, in confirmation question a Yes or No is given as an answer to an event expressed in the question.

Authors in [3] proposed six question categories which were tailored to general QA, namely: fact, list, reason, solution, definition and navigation. The proposed categories were motivated by work on users' classification goal proposed by Broder [18] and Rose and Levinson [19]. The given answer for the fact type of question will be a short phrase; this kind of question is asked to get a general fact as an answer. While, the given answer for list type of question will be a single phrase or a phrase with explanations or comments; this kind of question is asked to get a list of answers. Furthermore, the answer of for reason type of question should contain a variety of opinions or comprehensive explanations for a good answer summary in which sentence-level summarization can be employed; this kind of question is asked to get explanations or opinions as the answer.

In addition, the answer for solution type of questions usually formulate a sentence with a logical order. This kind of question is asked to solve a problem. While the definition type of questions are asked for the objective of getting a description of concepts as an answer in which this information usually could be found in Wikipedia. Finally, navigation type of questions are asked for the objective of finding websites or resources; sometimes the websites are given by name and the resources are given directly.

Furthermore, in [4] authors classified questions to 11 categories, which are: Advantage/Disadvantage, Cause and Effect, Comparison, Definition, Example, Explanation, Identification, List, Opinion, Rationale, and Significance. Advantages and disadvantages questions require certain number, while cause and effect questions explain the effect of something on something else. Furthermore, the answer of a comparison question usually outlines the differences and/or similarities between two or more entities. While, the answer of a definition question contains relatively short explanation or description.

In addition, the answer of an example question provides an example. On the contrary, the answer of an explanation question provides more explanation or more details than the what questions. Identification questions are type of questions that provide answers that allow the identification of something. Moreover, the list type of question provides a list of points which may or may not be in sequence. While, the answer of an opinion question contains 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.

In [1] questions were classified and labelled to six different categories, which are; causal, choice, confirmation (Yes-No questions), factoid (Wh-questions), hypothetical and list. This classification is based on the question types asked by the users and the answers given and categorization was motivated by the question types in English.

Yes-No questions or confirmation questions have an expected answer of either "Yes" or "No". Wh-questions or factoid questions contain any kind of information and any kind of information can be given as an answer or response. Furthermore, most factoid questions are formulated as an advice question, and are related to facts, current events, ideas and suggestions.

In addition, choice questions mainly offers choices in the question and usually contain two (or more) presented options. The objective of hypothetical questions is to have a general idea of a certain situation, it is mainly a what/if question. Finally, causal questions require further explanation as an answer. While, the answer of list questions takes the form of a list of entities or facts and plural terms are a highly reliable indicator of this type of question.

The most famous factoid question type taxonomy is the one of Li and Roth [5]. Many researchers focused on Li and Roth classification of question [10, 14, 16, 17, 8, 20, 9, 11, 12, 21, 22, 7]. Their two-layer taxonomy consists of a set of six coarse-grained categories which are abbreviation, entity, description, human, location and numeric value, 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. Table 2 shows the distribution of these classes, each coarse class contains a non-overlapping set of fine classes.

In the next sub-section we focus mainly on question classification using Li and Roth's categories [5] which is adopted in this study as it is the most widely used question intent taxonomy

Table 1: Summary of user intent categories for questions

Authors	Categories
Mohasseb, A, et at. [1]	Causal, Choice, Confirmation (Yes-No Questions), Factoid (Wh-Questions), Hypothetical and List.
Kolomiyets, O, et al. [2]	Hypothetical, Causal, Factoids, List, Definition, Relationship, Procedural, and Confirmation questions.
Bu, F, et al. [3]	Definition, Navigation, Fact, List, Reason and Solution,
Bullington, J, et al.[4]	Cause and Effect, Comparison, Definition, Advantage/Disadvantage, Example, Explanation, Identification, List, Opinion, Rationale and Significance.
Li and Roth [5]	Entity, Abbreviation, Description, Human, Location and Numeric as coarse classes; and Expression, Manner, Color, City.

in the literature.

2.2. Question Classification Methods

In this section we review related work about question classification methods and machine learning algorithms using Li and Roth's question categories.

Authors in [10] used compositive statistic and rule classifiers combined with different classifiers and multiple classifier combination methods. Moreover, many features such as dependency structure, Wordnet Synsets, Bag-of-Words, and Bigram were used with a number of kernel functions. In addition, an analysis has been conducted to identify the influence of using different ways of combining classifiers, such as Voting, adaboost, Artificial Neural Networks (ANN) and Transition-Based Learning (TBL), on the precision of the questions classification.

In [16] a method was proposed using feature selection algorithm to determine appropriate features corresponding to different question types. In addition, a new type of feature was designed, which is based on question patterns. A feature selection algorithm was applied to determine the most appropriate feature set for each question types. The proposed approach was tested using TREC dataset and SVM was used as the classification algorithm.

Authors in [14] proposed a statistical classifier which is based on SVM. The proposed classifier learns question word specific classifier by using prior knowledge about correlations between question words and types, i.e. a what question will be classified with SVM*what*.

Furthermore,a SVM-based approach for question classification was proposed in [17]. The proposed approach incorporates dependency relations and high-frequency words. Experimental result on the UIUC corpus showed that the introduced features can improve the baseline system significantly in which the accuracy has improved to 93.4% when the top word and dependency relation features are combined.

In [11] question classification method was proposed using three different classifiers, k-Nearest Neighbor (kNN), Nave Bayes (NB), and SVM. In addition, features such as using bag-of-words and bag-of-ngrams were used and a set of lexical, syntactic, and semantic features were also 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.

Moreover, authors in [9] used five machine learning algorithms which are, KNN, NB, Decision Tree (DT), Sparse Network of Winnows (SNoW), and SVM. In addition, two features were used; bag-of-words and bag-of-ngrams.

In [12] a head word feature was proposed and two approaches were presented to augment semantic features of such head words using WordNet. In addition, in this work authors adapted Lesks word sense disambiguation (WSD) algorithm and the depth of hypernym feature is optimized with further augment of other standard features such as unigrams. The proposed approaches reach the accuracy of 89.2% and 89% using linear SVM and Maximum Entropy (ME) models respectively over a standard benchmark dataset.

Moreover, authors in [8] proposed a compact feature set that uses typed dependencies as semantic features. The proposed feature set integrates only two simple dependencies of type nominal subject and prepositional object.

A hierarchical classifier was designed in [20], the proposed classifier takes in a training set and partitions it into a base train set to train the coarse classifier over all the possible question types. In addition, different classifiers has been used such as SVM, MaxEnt, NB and Decision Tree for primary and secondary classification. Results showed that the mix of a Maximum Entropy course classifier with a Naive Bayes fine classifier was the best combination in which results indicated that it was better to mix classifiers than to have the same type of classifier as both the primary and secondary.

In [21] authors used unlabeled questions in combination with labeled questions for semi-supervised learning. In addition, Tri-training were selected to improve the precision of question classification task in which two Tri-training methods were proposed; one that uses multiple algorithms for classifiers in Tri-training and the second is to use multiple algorithms for classifiers in combination with multiple views.

In addition, a two-level hierarchical classifier for question classification was proposed in [22]. The proposed classifier classifies the question sequentially two times by a coarse classifier and one of the six fine classifier. Moreover, different machine learning algorithms were used for the coarse classifier and fine classifiers such as supervised and semi-supervised learning.

Table 2: Li and Roth [5] questions classification

Coarse	Fine
ABBR	abbreviation, expression
DESC	definition, description, manner, reason
ENTY	animal, body, color, creation, currency, disease/medicine, event, food, instrument, language, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
HUM	description, group, individual, title
LOC	city, country, mountain, other, state
NUM	code, count, date, distance, money, order, other, percent, period, speed, temperature, size, weight

Finally, in [7] authors classified what-type questions by head noun tagging. In addition, different features such as local syntactic feature, semantic feature and category dependency were integrated among adjacent nouns with Conditional Random Fields (CRFs) to reduce the semantic ambiguities of head noun.

3. Proposed Approach

3.1. Factoid Questions Grammatical Features

This analysis was first introduced in [1]. Wh-questions (factoid) has its own characteristics, features, and structure that help in the identification and the classification process.

The main feature of a factoid question (Wh-Questions) is the presence of question words, this kind of question starts with a question word, such as *What, Where, Why, Who, Whose, When, Which*, e.g. *"What did the only repealed amendment to the U.S. Constitution deal with ?"*. In addition, this question could start with question words that do not start with "wh" such as *how, how many, how often, how far, how much, how long, how old*, e.g. *"How long does it take light to reach the Earth from the Sun ?"*

In addition, the structure of this type of question could begin with a Preposition followed by a question, *"P + QW"* such as *"In what year did Thatcher become prime minister?" OR "At what age did Rossini stop writing opera?"*. Also in many cases the question word could be found in the middle of the question, e.g. *"The corpus callosum is in what part of the body ?"*.

Most factoid questions are related to facts, current events, ideas and suggestions and could formulate an advice question, e.g. *"How do you make a paintball ?"*. In addition, some factoid questions could contain two types of question words, for example *"What does extended definition mean and how would one write a paper on it ?"*. Furthermore, factoid questions could have any kind of information given as an answer or response.

3.2. Question Classification Features

Three main features have been used for question analysis and classification which are, (1) Grammatical Features, (2) Domain specific Features and (3) Grammatical Pattern Features, these features transfer the question into domain specific grammatical pattern in which this new representation has the advantage of preserving the grammatical structure of the question.

3.2.1. Grammatical Features

Grammatical Features 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. Table 3. The grammatical features consist of the seven major word classes in English, which are Verb, Noun, Determiner, Adjective , Adverb , Preposition and Conjunction. In addition, we added a category for question words 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, Proper Nouns, Pronouns and Numeral Nouns as well as Verbs, such as Action Verbs, Linking Verbs and Auxiliary Verbs. In addition, it consists of other features such as singular and plural terms.

Table 3: Grammatical Features

Grammatical Feature	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
Common Noun	CN
Common Noun Other- Singular	CN_{OS}
Common Noun- Other- Plural	CN_{OP}
Question Words	QW
How	QW_{How}
What	QW_{What}
When	QW_{When}
Where	QW_{Where}
Who	QW_{Who}
Which	QW_{Which}

3.2.2. Domain Specific Grammatical Features

Domain-specific features (i.e. related to question-answering) where identified, which correspond to topics – these are listed in Table 4. Instead of further classifying the question to fine grained which is based on a large number of categories, we have used domain specific features to determine the type of question. For example, question type *ENTY* consists of fine grained categories such as religion, disease/medicine, event, product. These type could be identified using the following domain specific grammatical features: religion = religious terms PN_R, disease/medicine = health terms CN_{HLT} and PN_{HLT}, product = Products PN_P, event = events PN_E. Hence the domain specific grammatical features contain less categories than the fine grained categories proposed by Li and Roth but still could identify the different coarse categories.

Table 4: Domain Specific Grammatical Features

Domain specific Features	Abbreviation
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
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}

3.2.3. Grammatical Patterns

The question grammatical pattern help in the final identification of the question type, each factoid question type has a certain structure. For example, the following question which represent (HUM) type of question *"Who killed Gandhi?"* has the following grammatical pattern $QW_{Who} + AV + PN_C$. While, the question which represent (LOC) type of question *"What is the smallest country in Africa?"* has the following grammatical pattern $QW_{What} + LV + D + Adj + CN_{OS} + P + PN_G$. The different pattern representation help in distinguishing between different factoid question type.

A full description of how these features are used is provided in the following sections

3.3. Question Classification Framework

A Grammar-based framework for Questions Categorization and Classification (GQCC) is used [1]. The question classification framework takes into account the grammatical structure of the questions and combines grammatical features with domain-related information and grammatical patterns. The framework consists of three main phases; (1) Question Parsing and Tagging, (2) Pattern Formulation and (3) Question Classification. The following question from Li and Roth datasets will be used *"What causes asthma?"* to illustrate how these phases work.

(1) Question Parsing and Tagging: this step is mainly responsible for extracting users question terms. The system simply takes the question and parses to tag each term in the question to its terms' category. In this phase parsing the keywords and phrases is done by; first parsed compound words then single words. In addition, the term tagging is done by tagging each term to its grammar terminals; each term will be tagged to its highest level of abstraction (domain specific).

For the given example the question will be parsed and tagged as follow.:

Question: "What causes asthma?"

Terms extracted: What, causes, asthma

After parsing, each term in the question will be tagged to one of the terms category using tag-set the was proposed by [23] and [1]. The final tagging will be:

Question Terms Tagging: What=QW_{What}*, causes= AV, asthma=* CN_{HLT}

(2) Pattern Formulation: in this phase after parsing and tagging each term in the question, the pattern is formulated. This is done by matching the question with the most appropriate question pattern to help facilitate the classification processing and the identification of the factoid question type in the next phase.

For the given example, the following pattern will be formulated:

Question Pattern: $QW_{What} + AV + CN_{HLT}$

(3) Question Classification: This phase is done by using the patterns generated in Phase (2) by used using machine learning algorithms, 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 and a testing dataset. The training dataset is used for building the model, and the test dataset is used to evaluate the performance of the model.

For the given example, the question will be classified to the following question type.

Question Type: DESC

4. Experimental Study and Results

In the experimental study we investigate the ability of machine learning classifiers to distinguish between different question types based on grammatical features and question patterns.

Two machine learning algorithms, were used for question classification; Support Vector Machine (SVM) and J48. We used 1000, 2000 and 3000 questions that were selected from Li and Roth [1]. Their distribution is given in Table 5. Questions in the dataset are classified into two categories; coarse and fine, in this experiment coarse categories have been used.

Table 5: Data distribution

Question Type	1000	2000	3000
ABBR	18	30	45
DESC	211	419	655
ENTY	244	486	710
HUM	220	442	655
LOC	156	312	457
NUM	151	311	478

To assess the performance of proposed features and the machine learning classifiers two experiments have been conducted (1) using our proposed features using the Weka[2] software [24] and (2) using the most used features such as n-gram, Bag-of Words, Snowball Stemmer and stop word remover using Knime [3] software. The experiments were set up using the typical 10-fold cross validation. The results are presented in the next sub-section.

4.1. Results

In this section we present and analyse the results of the machine learning algorithms for each of the three set of questions. Table 6 shows the accuracy for GQCC and n-gram based classifiers for the 1000, 2000 and 3000 questions. In following sections the results will be discussed in more details.

Table 6: Accuracy of GQCC and n-gram based classifiers for 1000, 2000 and 3000 questions

Questions	**GQCC$_{SVM}$**	**n-gram$_{SVM}$**	**GQCC$_{J48}$**	**n-gram$_{J48}$**
1000	92.6%	87%	95.5%	86.7%
2000	95.1%	89.3%	96.6%	88.9%
3000	95.5%	92.4%	95.8%	91.1%

4.1.1. 1000 Questions

Table 7 presents the classification performance details (Precision, Recall and F-Measure) of the classifiers that have been used SVM and J48 using the proposed grammatical features, while Table 8 presents the classification performance details (Precision, Recall and F-Measure) of SVM and J48 using features such as n-grams, punctuation eraser, stop-word remover and snowball stemmer. Results show that Decision Tree (GQCC$_{J48}$) identified correctly (i.e. Recall) 95.5% of the questions and GQCC$_{SVM}$ identified correctly 92.6% of the questions when grammatical features were used while Decision Tree (n-gram$_{J48}$) identified correctly (i.e. Recall) 86.7% of the questions and n-gram$_{SVM}$ identified correctly 87.3% of the questions when features such as n-grams and snowball Stemmer were used.

Comparing the performance of the classifiers when 1000 questions were evaluated, GQCC$_{J48}$ had a better performance than GQCC$_{SVM}$, in which GQCC$_{J48}$ has the highest precision, recall and f-measure for all the classes and GQCC$_{SVM}$ has a similar precision (100%) as GQCC$_{J48}$.

When comparing the performance of n-gram$_{J48}$ and n-gram$_{SVM}$, n-gram$_{SVM}$ has a better precision, recall and f-measure for classes such as ABBR and NUM. In addition, both classifiers have similar precision, recall and f-measure for class type HUM. For class type DESC and ENTY n-gram$_{SVM}$ has a recall of (100%) while n-gram$_{J48}$ has better precision and f-measure.

Comparing the classification performance of GQCC$_{SVM}$ and n-gram$_{SVM}$. GQCC$_{SVM}$ has better precision and f-measure for class type DESC and ENTY while n-gram$_{SVM}$ has better recall (100%). For class type HUM and LOC, GQCC$_{SVM}$ has better Recall and n-gram$_{SVM}$ has better precision and f-measure. In addition, n-gram$_{SVM}$ has a (100%) precision, recall and f-measure for class type ABBR and higher precision, recall and f-measure than GQCC$_{SVM}$ for class type NUM.

Comparing GQCC$_{J48}$ and n-gram$_{J48}$, GQCC$_{J48}$ has a (100%) precision, recall and f-measure for class type ABBR, DESC and HUM. For class type ENTY, GQCC$_{J48}$ has higher recall and f-measure while n-gram$_{J48}$ has higher precision. While for class type LOC, GQCC$_{J48}$ has better precision and f-measure and n-gram$_{J48}$ has better Recall.

Table 7: Performance of the classifiers using grammatical features (1000 questions) - Best results are highlighted in bold.

	GQCC$_{SVM}$			**GQCC$_{J48}$**		
Class:	P	R	F	P	R	F
ABBR	**1.000**	0.833	0.909	**1.000**	**1.000**	**1.000**
DESC	0.995	0.995	0.995	**1.000**	**1.000**	**1.000**
ENTY	0.845	0.893	0.869	**0.873**	**0.955**	**0.912**
HUM	0.995	0.995	0.995	**1.000**	**1.000**	**1.000**
LOC	0.848	0.821	0.834	**0.936**	**0.840**	**0.885**
NUM	0.848	0.821	0.834	**0.986**	**0.940**	**0.963**

Table 8: Performance of the classifiers using n-gram features (1000 questions) - Best results are highlighted in bold.

	n-gram$_{SVM}$			**n-gram$_{J48}$**		
Class:	P	R	F	P	R	F
ABBR	**1.000**	**1.000**	**1.000**	**1.000**	0.800	0.889
DESC	0.887	**1.000**	0.940	**0.911**	0.984	**0.947**
ENTY	0.712	**1.000**	0.831	**0.965**	0.743	**0.840**
HUM	**1.000**	**0.788**	**0.881**	**1.000**	**0.788**	**0.881**
LOC	**1.000**	0.553	0.712	0.605	**0.978**	**0.748**
NUM	**1.000**	**0.933**	**0.966**	0.953	0.911	0.932

[1] http://cogcomp.org/Data/QA/QC/
[2] http://www.cs.waikato.ac.nz/ml/weka/
[3] https://www.knime.com

4.1.2. 2000 Questions

Table 9 presents the classification performance details (Precision, Recall and F-Measure) of the classifiers that have been used SVM and J48 using the proposed grammatical features, while Table 10 presents the classification performance details (Precision, Recall and F-Measure) of SVM and J48 using features such as n-grams, punctuation eraser, stop-word remover and snowball stemmer. Results show that Decision Tree ($GQCC_{J48}$) identified correctly (i.e. Recall) 96.6% of the questions and $GQCC_{SVM}$ identified correctly 95.1% of the questions when grammatical features were used while Decision Tree (n-$gram_{J48}$) identified correctly (i.e. Recall) 88.8% of the questions and n-$gram_{SVM}$ identified correctly 89.3% of the questions when features such as n-grams and snowball Stemmer were used.

When 2000 questions were evaluated, $GQCC_{J48}$ outperformed $GQCC_{SVM}$ in terms of precision, recall and f-measure for classes such as ABBR, ENTY, LOC and NUM. While, for classes such as DESC and HUM both classifiers had (100%) recall. For n-gram based classifiers n-$gram_{SVM}$ had better precision, recall and f-measure for classes such as ABBR, NUM. While, n-$gram_{J48}$ had better performance for class type ENTY. In addition, for class type DESC n-$gram_{SVM}$ had better recall while n-$gram_{J48}$ had better precision and f-measure. On the other hand, for class type LOC, n-$gram_{SVM}$ had better precision while n-$gram_{J48}$ had better recall and f-measure.

Comparing the performance of $GQCC_{SVM}$, $GQCC_{J48}$ and n-$gram_{SVM}$, n-$gram_{J48}$. $GQCC_{SVM}$ had a better precision, recall and f-measure for class type ABBR, DESC and HUM. while, n-$gram_{SVM}$ had better precision, recall and f-measure for class type NUM. Moreover, n-$gram_{SVM}$ has better precision and f-measure for class type ENTY. While, n-$gram_{SVM}$ has better recall. On the other hand, for class type LOC $GQCC_{SVM}$ has better recall and f-measure while n-$gram_{SVM}$ has higher precision.

Comparing $GQCC_{J48}$ and n-$gram_{J48}$, $GQCC_{J48}$ has better performance in terms of precision, recall and f-measure for classes such as ABBR, DESC, HUM and NUM. While for class type ENTY $GQCC_{J48}$ has higher recall and f-measure and n-$gram_{J48}$ has better precision. On the other hand, for class type LOC $GQCC_{J48}$ has better precision and f-measure while n-$gram_{J48}$ has higher recall.

Table 9: Performance of the classifiers using grammatical features (2000 questions)

	$GQCC_{SVM}$			**$GQCC_{J48}$**		
Class:	P	R	F	P	R	F
ABBR	**1.000**	0.967	0.983	**1.000**	**1.000**	**1.000**
DESC	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
ENTY	0.915	0.903	0.909	0.954	0.936	0.945
HUM	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
LOC	0.859	0.901	0.879	**0.881**	**0.929**	**0.905**
NUM	0.960	0.936	0.948	**0.977**	**0.952**	**0.964**

Table 10: Performance of the classifiers using n-gram features (2000 questions) - Best results are highlighted in bold.

	n-$gram_{SVM}$			**n-$gram_{J48}$**		
Class:	P	R	F	P	R	F
ABBR	**1.000**	**0.889**	**0.941**	**1.000**	**0.889**	**0.941**
DESC	0.920	**1.000**	0.958	0.933	**0.992**	**0.962**
ENTY	0.777	0.979	0.867	**0.959**	**0.795**	**0.869**
HUM	**0.973**	**0.827**	**0.894**	**0.973**	0.820	0.890
LOC	**0.896**	0.645	0.750	**0.650**	0.957	**0.774**
NUM	**0.978**	**0.957**	**0.967**	0.977	0.925	0.950

4.1.3. 3000 Questions

Table 11 presents the classification performance details (Precision, Recall and F-Measure) of the classifiers that have been used SVM and J48 using the proposed grammatical features, while Table 12 presents the classification performance details (Precision, Recall and F-Measure) of SVM and J48 using features such as n-grams, punctuation eraser, stop-word remover and snowball stemmer. Results show that Decision Tree ($GQCC_{J48}$) identified correctly (i.e. Recall) 95.8% of the questions and $GQCC_{SVM}$ identified correctly 95.5% of the questions when grammatical features were used while Decision Tree (n-$gram_{J48}$) identified correctly (i.e. Recall) 91.1% of the questions and n-$gram_{SVM}$ identified correctly 92.4% of the questions when features such as n-grams and snowball Stemmer were used.

When 2000 questions were evaluated, both $GQCC_{J48}$ and $GQCC_{SVM}$ had nearly similar performance, both classifiers had (100%) recall for class type ABBR and similar recall for classes such as ENTY and HUM. However, $GQCC_{J48}$ has higher precision and f-measure for these classes. In addition, $GQCC_{SVM}$ has better performance for LOC class while for classes such as DESC and NUM $GQCC_{SVM}$ has higher precision and $GQCC_{J48}$ has higher recall and f-measure.

Comparing the performance of n-$gram_{SVM}$ and n-$gram_{J48}$; n-$gram_{SVM}$ has higher performance for class type NUM while n-$gram_{J48}$ has higher performance (100%) precision, recall and f-measure for class type ABBR. For class such as HUM, LOC n-$gram_{SVM}$ has better precision and n-$gram_{J48}$ has better recall and f-measure. In addition, for class type DESC, n-$gram_{SVM}$ has better recall and f-measure while n-$gram_{J48}$ has better precision. while n-$gram_{SVM}$ has higher recall for class type ENTY and n-$gram_{J48}$ has higher precision and f-measure. On the other hand, for classes such as HUM and LOC, n-$gram_{SVM}$ has better precision and n-$gram_{J48}$ has better recall and f-measure.

Comparing the performance of $GQCC_{SVM}$, $GQCC_{J48}$ and n-$gram_{SVM}$, n-$gram_{J48}$. $GQCC_{SVM}$ has higher precision, recall and f-measure for class type ABBR and HUM. While, n-$gram_{SVM}$ has higher precision, recall and f-measure for class type NUM. In addition, for classes such as DESC, ENTY $GQCC_{SVM}$ has better precision and f-measure while n-$gram_{SVM}$ has better Recall. On the contrary, for class type LOC, $GQCC_{SVM}$ has better recall and f-measure while n-$gram_{SVM}$ has better precision.

Furthermore, $GQCC_{J48}$ has better performance for classes

such as DESC and HUM while, both classifiers ($GQCC_{J48}$ and n-$gram_{J48}$) have similar precision, recall and f-measure (100%) for class type ABBR. Moreover, for classes such as ENTY and NUM, $GQCC_{J48}$ has better recall and f-measure while n-$gram_{J48}$ has better precision and for class type LOC, n-$gram_{J48}$ has better Recall and $GQCC_{J48}$ has better precision and f-measure.

Table 11: Performance of the classifiers using grammatical features (3000 questions) - Best results are highlighted in bold.

	$\mathbf{GQCC}_{SVM}$			$\mathbf{GQCC}_{J48}$		
Class:	P	R	F	P	R	F
ABBR	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
DESC	**0.998**	0.998	0.998	**0.998**	**1.000**	**0.999**
ENTY	0.917	**0.920**	0.918	**0.937**	**0.920**	**0.928**
HUM	0.998	**0.998**	0.998	**1.000**	**0.998**	**0.999**
LOC	**0.861**	**0.908**	**0.884**	0.859	0.904	0.881
NUM	**0.987**	0.931	0.958	0.970	**0.948**	**0.959**

Table 12: Performance of the classifiers using n-gram features (3000 questions) - Best results are highlighted in bold.

	$\mathbf{n\text{-}gram}_{SVM}$			$\mathbf{n\text{-}gram}_{J48}$		
Class:	P	R	F	P	R	F
ABBR	**1.000**	0.909	0.952	**1.000**	**1.000**	**1.000**
DESC	0.961	**1.000**	**0.980**	**0.990**	0.954	0.972
ENTY	0.788	**0.981**	0.874	**0.977**	0.803	**0.881**
HUM	**0.995**	0.866	0.925	0.924	**0.934**	**0.929**
LOC	**0.922**	0.691	0.790	0.700	**0.971**	**0.813**
NUM	**0.991**	**0.942**	**0.966**	0.985	0.916	0.949

5. Discussion

These results indicate that in term of precision, recall and f-measure $GQCC_{J48}$ and $GQCC_{SVM}$ had the better performance when 1000, 2000 and 3000 questions were evaluated. In addition, for class type NUM, which consists of questions such as how many, how much and how long, n-$gram_{SVM}$ performed marginally better than both GQCC classifiers when 1000, 2000 and 3000 questions were evaluated, which indicate that n-gram based classifiers is more suitable in the identification of this type. While CQCC performed better for all other classes (ABBR, DESC, LOC, ENTY and HUM) in which combining grammatical features and domain specific grammatical features improved the classification of these type and enable the machine learning algorithms to better differentiate between different class types, since questions related to these type of classes contain terms related to companies name, geographical areas, places and buildings..etc. *(e.g "What does IBM stand for", "What is the name of the largest water conservancy project in China ?", "Who was Jean Nicolet ?"*

Comparing our approach with the state-of-the-art methods as shown in Table 13, the majority of the previous works used SVM for the classification process; in our experiments it has been shown that other classifiers like J48 could have a better performance and classification accuracy. The proposed hierarchical classifier in [5] classified questions into fine grained classes, using Sparse Network of Winnows (SNoW); using only syntactical features, the proposed approach achieved accuracy of 92.5% for coarse grained classes. In [9] bag-of-words features were used with different machine learning algorithms in which SVM performed better comparing with the other classifiers and has achieved an accuracy of 85.8% with coarse grained classes. Furthermore, In [12] head word features were used in addition to wordNet and unigrams; using liner SVM and maximum entropy models the proposed approach has achieved an accuracy accuracy of 89.2% and 89% respectively. In [14] the statistical classifier is based on SVM and has achieved an accuracy of 90.2% using coarse grained classes. In [7] head Noun tagging was used and was combined with syntactical and semantic features; for the classification process conditional random fields (CRFs) and SVM were used; the model achieved an accuracy of 85.6%.

In addition, in [21] the proposed method which is based on question patterns and designed features has achieved an accuracy of 95.2% using SVM. In [11] a combinations of semantic features with the lexico-syntactic features were used which achieved an accuracy of 96.2% for coarse classification. Work in [16] which was based on a new type of features and question patterns obtained an accuracy of 95.2% for coarse grain using SVM. Moreoever, The hierarchical classifier in [20] achieved accuracy of 77.8% using different classifiers such as SVM, MaxEnt, NB and Decision Tree. Finally, in [17] the proposed SVM-based approach achieved accuracy of 93.4% using a Bi-Gram mode and SVM kernel function.

In conclusion, GQCC had a better results than previous ones due to the ability of our approach to identify different classes of the factoid question using domain-specific information which facilitate the identification of domain categories, unlike previous works which used additional fifty fine-grained categories.

6. Conclusion and Future Work

In this paper, a Grammar-based framework for Questions Categorization and Classification (GQCC) was adapted. The framework make use of three main features which are, grammatical features, domain specific features and patterns. These features help in preserving the structure of the questions. In addition, the performance of different machine learning algorithms (J48 and SVM) were investigated for the classification of factoid questions. The results show that our solution led to a good performance in classifying questions compared with other state-of-arts approaches.

As future work, we aim to combine different features like semantic, syntactic and lexical attributes and compare the results. In addition, we aim to investigate the impact of handling class imbalance on the classification accuracy with approaches such as ensemble learning, since the labels distribution across the question datasets is imbalanced. We are also planning to test other machine learning algorithms to classify the questions.

Table 13: Previous approaches performance

Authors	Features	Algorithms	Accuracy
[5]	syntactical features	Sparse Network of Winnows (SNoW)	92.5%
[9]	bag-of-words features	SVM	85.8%
[12]	head word features, unigrams and wordNet	liner SVM/ Maximum entropy	89.2%/ 89%
[14]	syntactic and semantic features	SVM	90.2%
[7]	head Noun tagging, syntactical and semantic features	SVM	85.6%
[21]	question patterns and designed features	SVM	95.2%
[11]	semantic features with the lexico-syntactic features	KNN, NB, SVM	96.2%
[16]	question patterns	SVM	95.2%
[20]	part-of-speech, Parse Signatures and WordNet	SVM, MaxEnt, NB, Decision Tree	77.8%
[17]	Parts-of-Speech, Bi-Gram and Named Entities	SVM	93.4%

References

[1] A. Mohasseb, M. Bader-El-Den, M. Cocea, Question categorization and classification using grammar based approach, Information Processing & Management.

[2] O. Kolomiyets, M.-F. Moens, A survey on question answering technology from an information retrieval perspective, Information Sciences 181 (24) (2011) 5412–5434.

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

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

[5] X. Li, D. Roth, Learning question classifiers: the role of semantic information, Natural Language Engineering 12 (03) (2006) 229–249.

[6] D. Moldovan, M. Paşca, S. Harabagiu, M. Surdeanu, Performance issues and error analysis in an open-domain question answering system, ACM Transactions on Information Systems (TOIS) 21 (2) (2003) 133–154.

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

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

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

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

[11] M. Mishra, V. K. Mishra, H. Sharma, Question classification using semantic, syntactic and lexical features, International Journal of Web & Semantic Technology 4 (3) (2013) 39.

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

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

[14] D. Metzler, W. B. Croft, Analysis of statistical question classification for fact-based questions, Information Retrieval 8 (3) (2005) 481–504.

[15] T. Hao, W. Xie, F. Xu, 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, 2015, pp. 333–344.

[16] N. Van-Tu, L. Anh-Cuong, Improving question classification by feature extraction and selection, Indian Journal of Science and Technology 9 (17).

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

[18] A. Broder, A taxonomy of web search, ACM Sigir forum 36 (2) (2002) 3–10.

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

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

[21] T. T. Nguyen, L. M. Nguyen, A. Shimazu, Using semi-supervised learning for question classification, Vol. 3, Information and Media Technologies Editorial Board, 2008, pp. 112–130.

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

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

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: an update, ACM SIGKDD explorations newsletter 11 (1) (2009) 10–18.