Register Linear Based Model for Question Classification Using Costa Level Questions

SHANTHI PALANIAPPAN
Department Of Computer Applications
Sri Krishna College Of Engineering and Technology,
Kuniamuthur, Coimbatore
INDIA
shanthiphd2013@gmail.com

ILANGO KRISHNAMURTHI
Department Of Computer Science and Engineering
Sri Krishna College Of Engineering and Technology,
Kuniamuthur, Coimbatore
INDIA
ilangokrishnamurthi@gmail.com

Abstract: Question classification module of a Question Answering System plays a very important role in identifying and providing results according to the user expectations. Different methods are involved in the classification that can be applied to all kinds of domains like machine learning and lexical database. Identifying the relevant approach for question classification for a specific domain is one of the foremost tasks. A study on different levels of questions including Blooms taxonomy and Costa taxonomy made the researchers to focus more on different categories of questions. To overcome these issues, we employ a question classifier using Register Linear (RL) models for a specific domain. The Register Linear (RL) Classification Model classifies the complex questions in a linear manner where each input is assigned to only one class. The RL classification model identifies the role of semantics provided in the input space which is divided into decision regions with the decision surfaces to be of linear functions of input x (sentence) for different set of classes. Initially, the Register Linear model identifies the role of semantics in a sentence, and with these roles being identified, statistical relations between the concepts in the sentence are derived that produce a probability distribution over different set of classes. With these classifications, the exact answer type is identified. The model proposed gives better results in terms of execution time (time taken to categorize the queries), classification accuracy and result analyzing efficiency.

Keywords: Register Linear, Question Answering System, World Wide Web, Semantic Features, Statistical Information, Hierarchical Structure.

1 Introduction

Web pages retrieved by the search engines do not offer precise information and may hold irrelevant information even in top ranked results [1] that lead researchers to look for an alternate information retrieval system to provide answers for the user queries. Question Answering System is one of the information systems that is becoming more popular among different types of users for obtaining the information required. In such an answering system, question classification is important for efficient and fast information retrieval. In this work, a register linear model for question classification using Costa level questions is considered to build an effective classifier encompassing the semantic and syntactic information.

Various tools are used for identifying linguistics in different languages such as the WordNet to classify the text accurately. These tools help us to retrieve the relevant information from question answering systems that allow users to communicate with the system using any natural language. Question classification approach also uses text similarity method for pairs of snippets with semantic and statistical information by using lexical database [2] which lacks content to be presented in the core area being addressed in RL models using costal level questions. In [3], a novel algorithm called Topic-Sensitive PLSA is presented that extends the original probabilistic latent semantic analysis (PLSA) for identifying the semantic classification by introducing a small portion of information from the user with two types of constraints. This PLSA model does not concentrate on identifying higher level of questions, a basis being formed in RL model. The automatic question classification proposes a new type of questions and classifies these questions for better accuracy [1].

Recently the field of Natural Language Processing (NLP) requires an efficient algorithm and methodology to evaluate the similarity between short texts and sentences. To understand the natural language certain methods use expert system, which refers to irregular, complex, and diverse philosophical meaning in the context of human language [4, 5]. The flaw of the NLP is that it can be understood only with the help of human language which does not provide accuracy in processing complex queries. With the help of RL model, ontology process for Costa level II keywords improves the accuracy in processing complex queries.

An automatic readability index for the Arabic language was presented in [19] to facilitate the usage of readability index of Arabic language with the help of clustering analysis and support vector machine. Almost all systems incorporate question classification component which involves a set of rules to recognize small number of question types. They suffer from the inadequate coverage of rules and their incapability to simplify unused types of questions. There are few data sets needed for training machine learning approaches to question classification. In this paper, we propose a Register Linear (RL) model to build an effective classifier and apply flat parse representation for classification using Costa level II keywords [9].

Costa level is framed to assist the learners to be familiar with levels of questioning and in formulating and identifying higher levels of questions. Proposed RL model is used for complex questions to analyze, categorize, explain, classify, compare, contrast, infer, organize, and sequence the questions. Then with the help of Labeled Entity (LE) recognition, rich features are produced for higher performance. Then the RL is used which has been applied to Natural Language Processing (NLP) questions requiring complex and overlapping features. Based on the results of the classifier and features obtained using LE, the accuracy level is proved for complex queries. Finally the relevant answer type is identified based on the coarse class and the fine class.

The key idea is that the proposed Register Linear model incorporates syntactic and semantic information extracted from the questions. Finally, through experimental evaluation, using Costa level II keywords, the proposed method is shown to deliver excellent performance in terms of classification accuracy, execution time and result efficiency. The main contribution of this work is not only a linear classification model to build an effective classifier but also provide a flat representation to apply the rules for classification. The proposed method also produces a semantic parse representation, which improves better accuracy of both parsing and question categorization. Meanwhile, Costa level II keywords in RL model compare two or more questions to improve the accuracy while processing complex queries.

This paper is organized as follows. First the register linear question classifier model with a neat architecture diagram is constructed followed by RL preliminaries and Flat Parse based question classifier to effectively classify the question model in Section 2. Section 3 presents the experimental settings to conduct the register linear question classifier. The register linear question classifier model is evaluated and discussed in Section 4 and concluded in Section 5.

2 Related work and Motivation

The current e-learning application focuses learners to educate at all levels to increase the model of online educational patterns and to maximize the use of online discussion forums. Question answering system is one of the major parts of e-learning application to provide effective answers to the users' questions. We have two major levels of Question answering system like question classification and answer retrieval. This paper gives the first level of QA system.

There are many machine learning algorithms, manual algorithms and semantic based algorithms for classification. These entire algorithms depend on the domain used in the learning process. So, each of these algorithms obtain its own positive and negative results based on the domain knowledge. In [1], [6], [8], [10], [11], the datasets are collected from the TREC dataset, which is an open domain question dataset. In our approach, we focus on a specific domain dataset. These data are represented in the form of ontology for any domain. A new domain needs an essential classification to deal with the answer type [10]. During classification, the study of hierarchical classifiers can reduce the number of fine grain classes of the answer type [8] that made us propose a RL classifier model.

Question classification also uses the levels of questions to categorize the given questions. In [5], multiple 5W questions are only used for answering, which lacks in selecting the predefined taxonomies. Taxonomies like Blooms and Costa provide the levels of questions based on the questions. Level one question, makes users to recollect the information. Level two enables users to process the information. Level three requires users to think beyond the questions. Nowadays, users post their questions in many different forms like What, When, Where, How, Why, Which but some questions are given by the users in its own form like Identify, Compare, How would you compare..?, How would you contrast…?, State.., Give the names.., Can you explain, etc.,. In such cases, users do not follow the exact patterns of the questions. Many question classification uses only What, When, Where… types of questions and not the users' own pattern. To overcome all the question types and also to solve this issue, a novel RL classifier model is used based on the Costa level of questions [9].

In [11], question classification deals with the head word feature of the questions. A Word Sense Disambiguation (WSD) method is used for representing the hypernym of the head words. Question classification is the first and important phase which classifies the user questions. It also derives the expected answer types by extracting keywords and reformulates the questions into semantically equivalent multiple questions for complex queries [14].

Semantic similarity is the essential concept used in various fields such as artificial intelligence, natural language processing, information retrieval, relation extraction, document clustering and automatic data extraction. A novel semantic similarity method is also used in [4], the expert systems applications. So, it is essential for nay classification to use the semantic similarity in any applications [15] and high-frequency keywords are used for classification using semantic similarity concepts as discussed in [12] which is a time consuming process. On the other hand, with the separation of syntactic pattern and classification in RL model, the execution time to categorize the queries is optimal.

Almost all research on question classification uses semantic and syntactic analysis without using n-gram methods and bag-of-words methods. In order to get better accuracy, it has been identified that semantic features of questions plays an important role. Multi-lingual question answering system has also proved that greater accuracy is achieved by combining syntactic and semantic features [21]. Question classification also uses some machine learning algorithms for identifying the similarity between the questions [24], which could enhance the system in searching the answer easily. Typed dependencies is also extracted automatically from the dependency parser of the questions to increase the accuracy of the question classifications, for some datasets with 8.0% accuracy, in our paper we have tried to focus on the accuracy using semantic approach [25].

According to the classification techniques and the efficient answer type, exact results can be obtained [16], the Costa level II keywords provide efficient answers in RL model. The answer extraction is the final module to retrieve the answer for a question posted by the user. In order to provide the next-generation question answering systems with better querying support, a new method like RL method is needed to identify whether the questions are incoherent and therefore could not be answered [13].

Ontologies play an important role in representing the formal knowledge in a QA system, which increases the need for formal representation of concepts and relationships. These systems use semantic roles to extract accurate answer type [17] and [18], a clustering method is presented for modeling the similarity between set of concepts using entropy based methods. Compositional question answering systems [7] use divide and conquer approach for answer retrieval for some specific domain. In RL model, we use the geographical concept in the form of ontology for classifications. This paper explores analyzed techniques which combine them to provide a question classifier which considerably outperforms the previous state-of-the-art systems on the complex question classification test set. The RL approach also produces a probability distribution over classes when compared to other approaches.

The probability for each class represents the certainty of the decision whereas CoQUOS system [12] uses random walk algorithm for propagating the queries which uses only single class whereas the proposed RL approach uses multiple classes in parallel way and uncertainty is incorporated. Register linear question classification model using semantic and syntactic information, initially uses the WordNet database for categorizing the questions. Register Linear (RL) models improves these WordNet database systems by providing a probability distribution output for class labels with the ontology class using Costa level questions. Syntactic pattern constructed using the six forms of Costa level II keywords namely compare, separate, identity, and shows, analyze, and categorize. The probability distribution is then used by the question categorizing system as part of the final answer ranking. The question classification module parses the user question to identify the expected answer type [1].

The main aim of our paper is to follow up our future work on answering of the classified questions over the linked data to determine the answers. We would also enhance our system in identifying the similarity of questions after the classification mechanism [22].

A question answering system, lacks in answering the complex questions efficiently, which will involve combining several semantic relations in order to answer them [23], but any system could improve their accuracy when the questions are classified more efficiently even in answering the complex questions.

3 Register linear question classifier model

In RL model, each complex question clause constructs syntactical pattern based on Costa level questions, which is a flat parse representation that identifies the main verb and the other main categories of the complex question clause. As a complex sentence has subordinate clauses and usually has more than one syntactic pattern per sentence, each such pattern includes evaluation and is processed individually. Certain research on complex question clause uses a public parser that is suitable in the majority of cases, but there are certain sentences where the correct set of role fillers are not identified using the parse tree.

Each noun in owl data format is a member of more than one class, and therefore the list of its possible semantic frames is a combination of the semantic frames defined in each of the classes in which it participates and provides statistical information. The pattern is realized in the form of Costa level questions with the ontology design pattern consisting of metadata, pointers between questions and connections between the questions. It extracts all the semantic frames in a class and considers them to be possible semantic frames for each of the nouns that are members of class. Each noun class in RL also identifies a list of selection constraints for the semantic roles. For example, compare and contrast river and sea.

River[+water OR +regions] Vs Sea [+sea water OR +fresh water OR +lakes]

Rivers→ runsthrough → Louisana
State →population →>1000000

In the above example, the comparison of river and sea can be compared with different ways based on their key features like water, region etc, In the second example, the answer for the query "Identify the river that runs through Lousiana" and the third example provides the result of the query with the state having a population of less than 1000000 is addressed. These types of questions are considered in our work for classifying questions.

The architecture diagram of RL classification model is shown in the figure 1. The diagram describes the Register Linear (RL) Classification Model with ontology used by the Costa level questions. The Register Linear (RL) Classification Model follows the subsequent steps to effectively classify the complex questions. The questions initially follow the syntactic pattern construction. The syntactical pattern is constructed, in which each object is represented by a variable cardinality set of symbolic, nominal features. It represents pattern structures, captivating into account more compound interrelationships between the attributes than it is possible in the case of flat, numerical feature vectors of fixed dimensionality that are used in statistical classification. The pattern constructed in relationship with the question tags, (i.e., what, when, where, why) forms, the syntactic pattern constructed that gives us the lexical constraints like Noun{NN}, Verb{VV} and other terms. The noun and verb keywords are analyzed with the semantic meaning using WordNet. Question analysis extracts all the useful information from the question to obtain a set of relevant classes in the ontology. Flat parse representation retrieves the queries related to the Costa level II keywords to obtain only the relevant information from the given ontology.

Subsequently, Register Linear Classification is illustrated using all the classification based on the classes. The RL model categorizes the classes from the set of questions. After the RL classification, entity recognition is performed on the labels for effective complex question classifier. Register Linear Classification are processed using the Labeled Entity Recognition for effective complex question classifier. For instance, labeled entity recognition is performed on question tags to list the classification. This type of recognition is used to fetch the effective result for the complex queries (i.e. two or more questions).

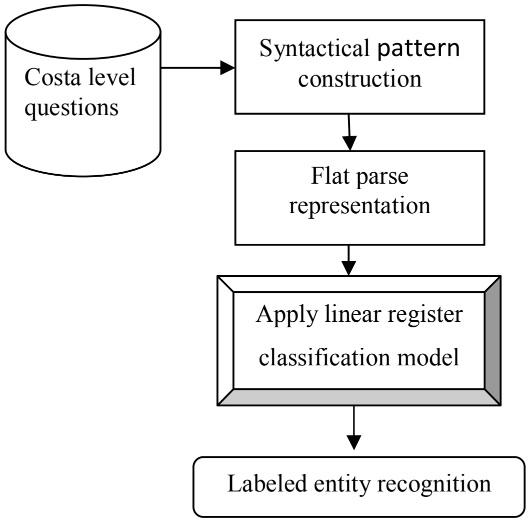

Costa level questions
Syntactical pattern construction
Flat parse representation
Apply linear register classification model
Labeled entity recognition

Fig.1: Architecture Diagram of Register Linear (RL) Classification Model

3.1 RL preliminaries

Register Linear models produce a probability distribution over multiple classes and have the advantage of handling large numbers of complex overlapping features through ontology patterns. These RL models have the following form for deriving the complex question pattern,

$$m(y|x,\sigma) = \frac{1}{C(x|\sigma)}\exp(\textstyle\sum_{i=1}^{n}\sigma_i f_i(x,y)) \qquad (1)$$

Where, i = 1,2,..,n with f_i are weights of the observation 'x' and the class label y. σ_i are the model parameters, and C(x | σ) is the normalization function. The feature weights correspond to the contribution that each complex feature should make towards a question classification with a large, complex positive feature weight indicates a feature that is strongly associated with a particular class. Feature weights may also be negative, in which case they indicate a disassociation.

The feature functions $f_i$are arbitrary functions of an observation and its label. Usually in RL classification these are binary valued functions that are defined for each observation. The RL model feature functions are shaped by the combination of the class label and predicate features that are highly required for ontology learning; for example:

$$f_i(x,y) = \begin{cases} 1\ ifwordinxandy = comparemanner \\ 0\ otherwise \end{cases} \qquad (2)$$

For example, consider the following two queries given below:
Identify the highest peak in the world
How will you compare road 90 and road 80?

For the above two queries, the parameters of questions are

Highest peak → world
Road 90 → road 80

The complex query feature would have the value 1 if the complex question x contained the word and the class y was 'Compare' manner. If these conditions were not met, then the complex feature would be inactive (0). The additional complex query features are functions of the class hierarchy in order to assemble a hierarchical classifier for effective classification. From RL model it predicts the most consistent probability allotment from the set of model and satisfies the constraints in the training data.

RL model fits the training data and prefers the smoothest one. Instead of solving the constrained optimization problem for the complex questions, it equivalently finds the utmost likelihood approximate for the feature weights (σ) given in the training data. In order to derive values for the model parameters, the likelihood of the training observations are maximized using maximum likelihood objective function, which estimates the parameters of RL model. The maximum likelihood of RL model uses an objective function f(.) with the matrix to be stored is denoted by ∑ with the variance provided using σ. However, maximum likelihood parameters for register linear models have a tendency to over fit the data.

$$m_0(\sigma_i) = \frac{1}{\sqrt{2\pi\tau^2}}\exp(-\frac{\sigma_i^2}{2\tau^2}) \qquad (3)$$

Consequently the maximum likelihood objective function for the model parameters is:

$$L = \textstyle\sum_k log_m(y^k\,|x^k,\sigma) + \sum_{i=1}^{n} logm_0\,(\sigma_i) \qquad (4)$$

$$= \textstyle\sum_k(\sum_{i=1}^{n}\sigma_i f_i\,(x^k,y^k) - \log\frac{1}{C(x^k|\sigma)}) \qquad (5)$$

$$-\textstyle\sum_{i=1}^{n}\frac{\sigma_i^2}{2\tau^2} + \text{Constant} \qquad (6)$$

Where x^k, y^k are the k^{th} training observation and its label respectively for the observations x^1, x^2, x^3 to x^k in order to train the RL model must maximize Eqn 4, which achieved using unconstrained minimization techniques. The vertical bar "|" denotes the separation between the two labels x^k, y^k applied to objective function f(.). Each iteration of algorithm, the objective function and the gradient are being calculated; therefore the complexity of the training algorithm is linear based on the number of training observations and features.

3.2 Flat parse based query classifier

Representation of complex questions differs from demonstration of declarative sentences and deserves special attention. For complex sentences representing questions, classification of statement is done in the same way that communicates to the question in a simple way and then in a similar way declarative complex sentences are constructed for ontology entities. The more similar the queries the more they denote the concept in a similar manner. Figure 2 represents the flat parse representation for Question classifier.

The Register Linear model classifier classify the 6 classes of ontology based on Costa level II keyword and compare the queries using the semantic interpretation of ontology and statistical information with flat parse representation. The result of parsing is a dependency indicating how the noun in the complex queries interacts syntactically. Using question classifier, class 1 uses the separate and compares keywords for question classifier and in separate keyword it separates the highest and lowest point of

Colorado. Class 2 uses to classify and categorize for querying the word.

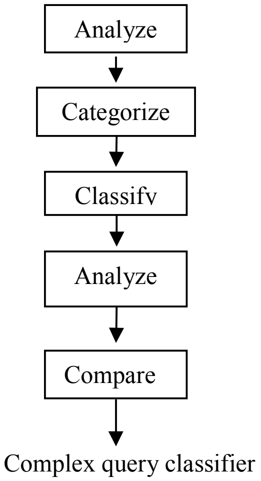


Fig. 2: Flat Parse Representations for Query Classifier

For instance, it categorizes the River based on length which runs through New Mexico and Categorize mountains based on height which is placed in Alaska? This question uses the "what, which" question tags to categorize the results. Figure 3 summarizes the entire process involved in register linear question based classifier model using Costa level questions.

Table 1: Entity type and the Class type

COARSE CLASS	FINE CLASS
ABBREVIATION	abbreviation, expression abbreviated
ENTITY	animal, body, color, creative, currency, diseases and medical, event, food, instrument, Lang, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word , name
DESCRIPTION	definition, description, manner, reason, comparison, analysis
HUMAN	group, title, description
LOCATION	city, street, river, country, mountain, other, state
NUMERIC	code, count, date, distance, money, order, other, period, percent, speed, temp, size, weight, points, population(count), weather, low point, high point,
IMAGE/ SPATIAL FEATURES	Area (Polygon, lines, volumes, grids, points)

Figure 3 shows the entire process involved in register linear question based classifier using Costa level questions for efficient classification and to improve the accuracy in processing complex queries. Initially, Costa level questions are given as input, followed by syntactical pattern being constructed with the help of six forms of questions namely compare, separate, identity, show, analyze, and categorize. Next, flat representation form is provided to identify the main verb and to obtain other main categories for question classifier. Followed by this, register linear classification is applied to derive complex query pattern and finally, labeled entity recognition is used to fetch the result for complex queries involving two or more questions for the purpose of effective complex question classifier.

The question classifier identifies the question types along with the answer type of the given questions. According to our data in the ontology, each question is analyzed and classified based on the questions. Later, the terms in Nouns and Verbs are extracted and semantically checked to identify the coarse class and the fine class of the given questions to represent its relevant answer type as given in table 1.

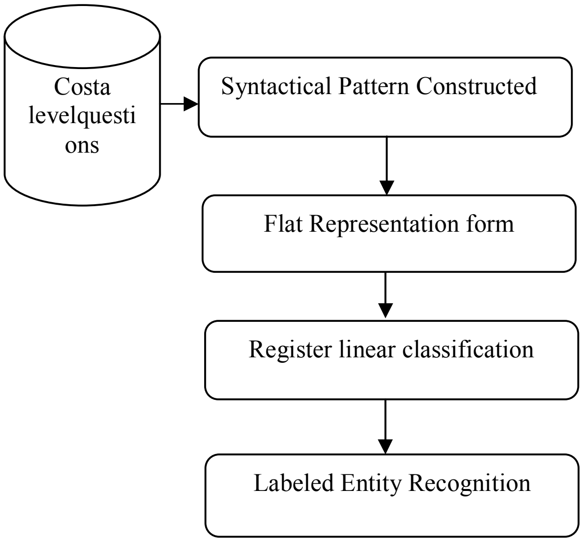

Costa levelquestions
Syntactical Pattern Constructed
Flat Representation form
Register linear classification
Labeled Entity Recognition

Fig. 3:Entire process of register linear question based classifier

3.3 Algorithm for RL model classification

The RL classification algorithm compares the sentence pattern against each of the possible semantic role frames extracted from WordNet. It compares the constituents before and after the verb in the sentence.

3.3.1 Algorithm for question classification

The algorithmic description for classification of question is given below:

```
Input: Question
Output: Complex Question Classifier
Begin
Step 1: Construction of Syntactic pattern
//in relationship with the question tags, (i.e., what, when, where, why, how) or
//in relationship with the question tags (i.e., identify, compare, analyze, show, separate…) forms
Step 2: Represent Flat parse form
//Register Linear Classification method
Step 3: Derive Complex Query pattern using eqn (1)
// Complex Question Classifier based on Labeled Entity Recognition using class label and predicate features
Step 4: Complex question classifier using eqn (2)
Step 5: If $f_i(x, y) =1$ then x contains the word
Step 6: Else If $f_i(x , y) =0$ then "Invalid"
Step 7: Measure the complex question classifier based on weight using eqn (3)
Step 8: Returns first set of function assignments as a final result (i.e. appropriate class for the given question)
Step 9: Find out the relevant answer type
End
```

3.3.2 Algorithm for identifying the relevant answer type

The algorithmic description for identification of relevant answer type is given below:

```
Input: Question, type_of_class
Output: Answer type
Begin
Step 1: if type_of_class = true then
        Extract Noun, Verb from question
        Find the synonyms of Noun, Verb
        //using Wordnet
Step 2: identify the coarse class (may be entity type) for the noun and verb
Step 3: find out the fine class for the given Noun/Verb term
Step 4: return the relevant answer type
```

The first step involved in the RL model classification is the construction of syntactic pattern with the help of six forms of Costa level II keywords namely compare, separate, identify, and shows, analyze, and categorize. With more complexity involved in each sentence having subordinate clauses, with more than one syntactic pattern, each such pattern is processed individually in relationship with the question tags, (i.e., what, when, where, why, how forms). The Question analysis extracts all the useful information from the question to form a set of relevant information.

The second step involved in the RL model classification is the representation of flat parse form that identifies the main verb and to obtain other main categories for question classifier. The task of flat parse representation is to retrieve the queries related to the Costa level II keywords and to extract relevant information. The third step involved is the actual register linear classification to derive complex question pattern and produce a probability distribution over multiple classes. Multiple results are identified when there are two or more phrases in a complex query. They are possible semantic role realizations, if there are two or more semantic frames for which matches were found.

The final step is the labeled entity recognition that is used to fetch the result for the complex queries involving two or more questions for effective complex question classifier. The complex question classifier will have the value 1 if the complex question x contained the word and the class 'y' was 'Compare' manner. If these conditions were not met, then the complex query classifier would be inactive (0) as stated earlier. To select the correct function assignment, weighting function is used that allocate scores to each consequence and returns the one with the highest score. For each identified role the weighting function adds one point if the role does not have any selection restrictions and two points if there are restrictions. The total score for a RL solution is the sum of the scores for each identified roles. The solution with the highest score is selected for effective query classifying.

4 Results on register linear question classifier

Extensive experiments conducted with various conditions using JAVA platform also included ontology in order to analyze the different models. Initially, question classification takes place using the WordNet dataset. WordNet is an English lexical database containing about 120000 entries of nouns, verbs, adjectives and adverbs, hierarchically organized. They are considered as ontology that can be applied in various question answering tasks in synonym groups linked with relations such as hypernym, hyponym, holonym and others. The experiments with WordNet designate that use of semantic information for question classification seriously recover the performance of Question Answering Systems. All of the works discussed above reported very high precision for question classification.

The owl data comprises sample ontological geographical[1] data sets, each supply a information base in OWL and English questions. Experiments conducted using owl based dataset is based on six classes, shows that the time taken to categorize the queries is less. Moreover the questions being classified achieve higher classification accuracy using efficient question categorizing systems.

Register linear question classification model using semantic and syntactic information uses the geographic data, consisting of more than 877 English geographic questions. Some of the geographic data questions are (i) Identify which state has more lakes, (ii)Separate city and capital of Arizona, (iii) Compare road 80 and road 90, (iv) Show roads that passes through Newyork, For each English question, there is also a consequent logical representation stated as Prolog terms. This is illustrated in the Appendix A.

5 Performance comparison of register linear question classifier

RL question classifier model is measured in terms of execution time, classification accuracy and overall result analysis of the system. With the ontology in hand, the Costa level II keywords are used to compare the more complex queries. Level two questions are used to process information and retrieve the effective result on the overall system using the RL model. In our RL model, ontology design patterns are constructed with the help of Costa level questions and comparison is made against the short text modeling method to compare the different parametric values. Execution time on RL model is the amount of time it takes to categorize the queries, in terms of seconds (sec).

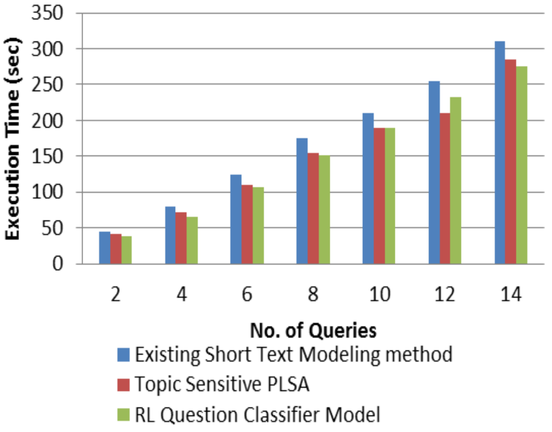
Accuracy in classification depends on the amount of effective result obtained based on the Costa level II keywords. Classification accuracy based on keywords measure (in terms of percentage) the exact categorization of complex queries. Result efficiency on overall system is defined as the improved percentage in classifying the complex owl geographic queries based on 6 classes of keywords. The execution time is measured and compared with the existing short text modeling method and PLSA method and our RL question classifier outperforms based on the questions extracted from the owl data form.

Table 2: Tabulation of Execution time

No. of queries	Execution Time (sec)		
	Existing Short Text Modeling method	Topic Sensitive PLSA	RL Question Classifier Model
2	45	42	38
4	80	72	66
6	125	110	107
8	175	155	152
10	210	190	189
12	255	210	232
14	310	285	275

Figure 4 describes the time taken to categorize the questions on the owl based geographic data. Compared to the existing Short Text Modeling method [2], probabilistic latent semantic analysis (PLSA) [3], the proposed question classifier achieves the classification with lesser time. This is because, RL Question Classifier Model with ontology entities fit the training data in such a way that it prefers the smoothest data and results in minimizing the time taken to categorize the queries whereas the PLSA model present an objective function that tradeoffs the likelihood of observed data and enforcement of constraints resulting in comparatively higher execution time. Additionally, the constrained optimization problem for the complex questions at the same time finds the likelihood approximation for the feature weights (σ) from the owl data queries. The variance achieved using RL Question Classifier Model is 10 – 20 % reduced when compared with the short text modeling method and 5-15% reduced when compared to the PLSA model.

Fig. 4: Measure of Execution time



[1] *https://files.ifi.uzh.ch/ddis/oldweb/ddis/research/talking.../owl-test-data*

Table 3: Tabulation of Classification Accuracy

Keyword Class	Classification Accuracy (%)		
	Existing Short Text Modeling method	Topic Sensitive PLSA	RL Question Classifier Model
1-Compare	85	87	90
2- Separate	88	90	91
3- Identify	91	92	95
4- Shows	88	90	91
5- Analyze	89	91	92
6- Categorize	90	92	93

The classification accuracy of the queries is based on the Costa II keywords with 6 classes. The classification accuracy is tabulated in Table 4. The value of the proposed RL query classifier is compared with the existing Short Text Modeling method and PLSA model. Figure 5 describes the classification accuracy based on the keyword class. Compared to the existing Short Text Modeling method [2], PLSA[3] model, the proposed RL query classifier is 2-5 % improved for performing classification process when compared to short text modeling and 3% improved when compared to PLSA.

As illustrated in figure 5, the classification accuracy of RL model with ontology fragment is higher because of the functions of the class hierarchy that generates a hierarchical classifier resulting in effective classification. The complex query feature has the value ‘1’ if the complex question ‘x’ shows the word and the class ‘y’ is ‘compare’ form. If the condition is met, then it categorizes that form of complex query effectively. If these conditions are not met, then the complex feature would be inactive to (0).

Table 4 Tabulation of Overall Result Analysis

No. of instance	Overall Result Analyzing Efficiency (%)		
	Existing Short Text Modeling method	Topic Sensitive PLSA	RL Question Classifier Model
5	76	77	90
10	79	82	92
15	80	85	92
20	81	87	92
25	82	90	93
30	84	91	95
35	85	92	93
40	88	94	98

Figure 6 describes the overall result ratio of the RL model and Short Text Modeling method [2] and PLSA [3]. The RL model improves the result by 8 – 15 % when compared to the short text modeling method and 3-5% when compared to the PLSA model because the PLSA model uses only two constraints namely must-link and cannot-link constraints. The instances taken for the evaluation varies from 5, 10…40. Existing Short Text Modeling method is not effective with the set of classes for combining semantic and statistical information.

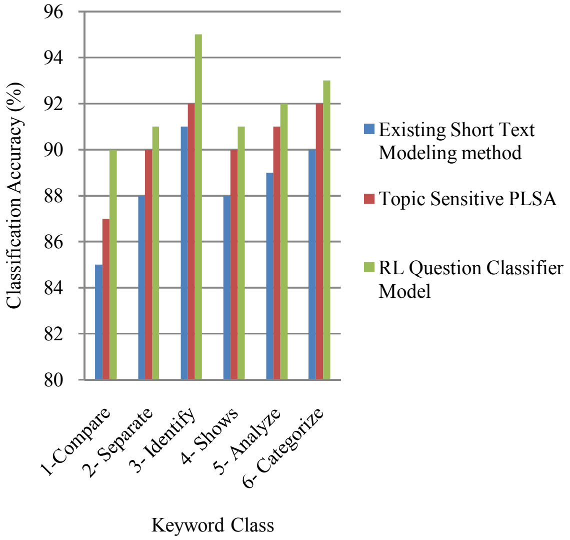


Fig. 5: Measure of Classification Accuracy

The overall result analyzed is effective because, instead of solving the constrained optimization problem for the complex questions, the RL model combines the semantic and syntactical information and finds the maximum likelihood approximate for the features weights (σ) for the given training data.

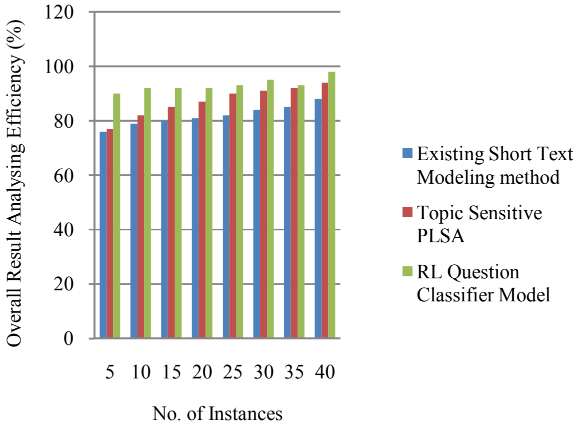


Fig. 6: Measure of Overall Result

Finally, it is being observed that the question classifier in combination with a log-linear model, obtain better results. Each instance uses the RL algorithm with the objective function; therefore the complexity of the training algorithm is linear with improved result percentage on the owl geographic data questions. It immediately results in an improvement in the accuracy and efficiency of question categorizing systems.

6 Conclusion

Register linear question classification model using semantic and syntactic information is our major task used to build an effective classifier with the ontology dataset using Costa level II keywords. Initially, RL constructs the syntactical pattern between the concepts in the sentence. Flat parse representation is the next step used for applying the rules for classification. The parser produces a semantic parse representation, which vastly improves the accuracy of both parsing and question categorization. RL model produces a probability distribution over a set of 6 class keywords. The ontology process for Costa level II keywords in RL model compare two or more questions to improve the accuracy in processing complex queries. A set of semantic features and statistical information are calculated and the approach makes use of the question classes in an effective manner. The result is a question classifier that outperforms previous short text modeling method and probabilistic latent semantic analysis (PLSA), in terms of execution time, classification accuracy and result analyzing efficiency. The experimental result of RL model attains effective result efficiency on the overall system. Approximately 3.5 % improvement in the accuracy is shown using RL for classifying the complex questions. Our work can be further improved by using enhanced semantic analysis techniques. In our future work, based on the identified classifications, we would enhance our system for answering the questions consuming minimum time.

Appendix A

List of sample complex questions:

1	Identify the state that has the most lakes?
2	Compare road 90 and road 80?
3	Show the road that passes through Newyork?
4	Identify the state that has abbreviation al?
5	Identify the city named city of Newyork?
6	Identify the capital city of Newyork?
7	Separate road no 95 and road no 90?
8	Identify the place that has the highest point of Arkansas?
9	Identify the place that is lowest point of Alaska?
10	Separate highest and lowest points of Colorado?
11	Identify the river that runs through Louisiana?
12	Identify the state having the population less than 1000000?
13	Separate city and capital of Arizona?
14	Identify the state that has population more than 1000000?
15	Identify the state that has the lowest point of elevation?
16	Categorize rivers based on length which runs through Newmexico?
17	Identify the state that has the tallest mountain?
18	Identify the state that has the longest straight road?
19	Identify the state that has the fewest neighboring state?
20	Identify which state has the most neighboring state?
21	Show rivers which passes through Arizona?
22	Analyse the state of New york?
23	Analyse the river of Delaware?
24	Categorize mountains based on height which placed in Alaska?
25	How will you compare road 90 and road 80?
26	What are all the factors to compare river and lakes?
27	Can you identify the state that has more lakes?

References:

[1] Santosh Kumar Ray., Shailendra Singh., B.P. Joshi., "*A semantic approach for question classification using WordNet and Wikipedia*," Pattern Recognition Letters., Elsevier Journal., pp. 1936-1938, 2010

[2] Liu Wenyin , Xiaojun Quan., Min Feng, Bite Qiu., "*A short text modeling method combining semantic and statistical information*," Information Sciences., Elsevier Journal., pp. 4033-4037, 2010

[3] Ke Zhou, Gui-Rong Xue, Qiang Yang and Yong Yu, "*Learning with Positive and Unlabeled Examples Using Topic-Sensitive PLSA*", IEEE Transactions On Knowledge and Data Engineering, Vol. 22, No. 1, pp. 4011-4014, January 2010

[4] Ming Che Lee., "*A novel sentence similarity measure for semantic-based expert systems*," Expert Systems with Applications, Elsevier Journal., pp. 6392-6395, 2011

[5] Sibel Yaman., Dilek Hakkani-Tur., Gokhan Tur., Ralph Grishman., Mary Harper., Kathleen R. McKeown., Adam Meyers., Kartavya Sharma., "*Classification-Based Strategies for Combining Multiple 5-W Question Answering Systems*," ISCA, pp. 2704-2708, 2009

[6] Hakan Sundblad, "*Question Classification in Question Answering Systems*", Linköping Studies in Science and Technology, Department of Computer and Information Science, Linköpings, university, SE-

581 83 Linköping, Sweden, Thesis No. 1320, pp. 5-8, 2007

[7] Hyo-Jung Oh., Ki-Youn Sung., Myung-Gil Jang., Sung Hyon Myaeng., "*Compositional question answering: A divide and conquer approach*," Information Processing and Management., Elsevier Journal., pp. 1022-1025, 2011

[8] Helena G_omez Adorno y David Pinto y Yuridiana Alem_an y Nahun Loya, "*A Question Classification study based on machine learning*", pp. 3-5, 2011

[9]www.nscsd.org/webpages/jennisullivan/files/levels%20of%20questioning%281%29.pdf, Costa level of questions.

[10] Lei SU, Zhengtao YU, Jianyi GUO, Yun LIAO, *"Domain Adaptation for Question Classification*", Journal of Computational Information Systems, pp. 3262-3267, 2011

[11] Zhiheng Huang, Marcus Thint, Zengchang Qin, "*Question Classification using HeadWords and their Hypernyms*", Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 927-936, 2008

[12] Zhengtao Yu, Lei Su c, Lina Li a, Quan Zhao a, Cunli Maoa, Jianyi Guo, "*Question classification based on co-training style semi-supervised learning*", Elsevier Journal, Pattern Recognition Letters, Volume 31, Issue 13, pp. 1975-1980, 2010.

[13] Ali Mohamed Nabil Allam and Mohamed Hassan Haggag, "*The Question Answering Systems: A Survey*", International Journal of Research and Reviews in Information Sciences (IJRRIS), Vol. 2, No. 3, pp. 211-220, 2012

[14] Poonam Gupta, Vishal Gupta, "*A Survey of Text Question Answering Techniques*", International Journal of Computer Applications, pp. 11-15, 2012

[15] Muthukrishnan Ramprasath, Shanmugasundaram Hariharan, "*Using Ontology for Measuring Semantic Similarity for Question Answering*", IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), pp. 7-11, 2012

[16] Deborah L. McGuinness, "*Question Answering on the Semantic Web*", Stanford University, IEEE Computer Society, Intelligent Systems, pp. 17-25, 2004

[17] Paloma Moreda., Hector Llorens., Estela Saquete., Manuel Palomar., "*Combining semantic information in question answering systems*," Information Processing and Management, Elsevier Journal., pp. 870-885, 2011

[18] Koskela, M, Helsinki, Smeaton, A.F, Laaksonen, J., "*Measuring Concept Similarities in Multimedia Ontologies: Analysis and Evaluations*", IEEE Transactions on Multimedia, Volume: 9, Issue: 5, pp.110-117, 2011

[19] Abdel-Karim Al-Tamimi, Manar Jaradat, Nuha Aljarrah, and Sahar Ghanem, "*ARI: Automatic Arabic Readability Index*", IAJIT, March 12, 2013.

[20] Seyed Sadatrasoul, Mohammad Gholamian, and Kamran Shahanaghi, "*Combination of Feature Selection and Optimized Fuzzy Apriori Rules: The Case of Credit Scoring*", IAJIT, December 28, 2013.

[21] Anbuselvan Sangodiah, Manoranjitham Muniandy and Lim EanHeng, "*Question Classification Using Statistical Approach: A Complete Review*", Journal of Theoretical and Applied Information Technology, Vol.71 No.3, pp 386-395, 2015.

[22] SaeedehShekarpour, Edgard Marx, Axel-CyrilleNgongaNgomo, and Sören Auer, "*SINA: Semantic Interpretation of User Queries for Question Answering on Interlinked Data*", Journal of Web Semantics Science, Services and Agents on the World Wide Web, 2014.

[23] DimitarHristovski, DejanDinevski, Andrej Kastrin and Thomas C Rindflesch, "*Biomedical question answering using semantic relations*", BioinformaticsKnowledge-based analysis, Volume 16:6, 2015.

[24] Quan Hung Tran, Minh Le Nguyen and Son Bao Pham, "*Question Analysis for a Community-Based Vietnamese Question Answering System*", Knowledge and Systems Engineering, Advances in Intelligent Systems and Computing, Volume 326, pp 641-651, 2015.

[25] Phuong Le-Hong, Xuan-HieuPhan and Tien-Dung Nguyen, "*Using Dependency Analysis to Improve Question Classification*", Knowledge and Systems Engineering, Advances in Intelligent Systems and Computing Volume 326, pp 653-665, 2015.