Undefined 0 (2017) 1–0
IOS Press

Difficulty-level Modeling of Ontology-based Factual Questions

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Vinu E.V * and P Sreenivasa Kumar
Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India
E-mail: {vinuev,psk}@cse.iitm.ac.in

Abstract. Semantics-based knowledge representations such as ontologies are found to be very useful in automatically generating meaningful factual questions. Determining the difficulty-level of these system generated questions is helpful to effectively utilize them in various educational and professional applications. The existing approaches for finding the difficulty-level of factual questions are very simple and are limited to a few basic principles. We propose a new methodology for this problem by considering an educational theory called Item Response Theory (IRT). In the IRT, knowledge proficiency of end users (learners) are considered for assigning difficulty-levels, because of the assumptions that a given question is perceived differently by learners of various proficiencies. We have done a detailed study on the features/factors of a question statement which could possibly determine its difficulty-level for three learner categories (experts, intermediates, and beginners). We formulate ontology-based metrics for the same. We then train three logistic regression models to predict the difficulty-level corresponding to the three learner categories. The output of these models is interpreted using the IRT to find the question's overall difficulty-level. The performance of the models based on cross-validation is found to be satisfactory and, the predicted difficulty-levels of questions (chosen from four domains) were found to be close to their actual difficulty-levels determined by domain experts. Comparison with the state-of-the-art method shows an improvement of 8.5% in correctly predicating the difficulty-levels of benchmark questions.

Keywords: Difficulty-level estimation, Item response theory, Question generation

1. Introduction

A considerable amount of effort has been invested into the creation of a semantics-based knowledge representations such as ontologies where information is formalized into machine-interpretable formats. Among these are SNOMED CT[1], BioPortal[2], Disease ontology[3], to name a few, which capture domain-specific knowledge. Given these knowledge repositories, the opportunity for creating automated systems which utilize the underlying knowledge is enormous. Making use of the semantics of the information, such systems could perform various intelligently challenging operations. For example, a challenging task which often required in an e-Learning system is to generate questions about a given topic which match the

*Corresponding author. E-mail: vinuev@cse.iitm.ac.in, mvsquare1729@gmail.com
[1]http://www.snomed.org/
[2]http://bioportal.bioontology.org/
[3]http://www.berkeleybop.org/ontologies/doid.owl

end users' (learners') educational need and their proficiency level.

The problem of generating question items from ontologies has recently gained much attention in the computer science community [1,2,6,7,22,23]. This is mainly due to the utility of the generated questions in various educational and professional activities, such as learner assessments in e-Learning systems, quality control in human computational tasks and, fraud detection in crowdsourcing platforms [19], to name a few.

Traditionally, question generation (QG) approaches have largely focused on retrieving questions from raw text, databases and other non-semantics based data sources. However, since these sources do not capture the semantics of the domain of discourse, the generated questions cannot be machine-processed, making them less employable in many of the real-world applications. For example, questions that are generated from raw text are suitable only for language learning tasks [5]. Using semantics-based knowledge sources in QG has various advantages, such as (1) in ontologies, we model the semantic relationships between domain entities, which help in generating meaningful and machine-processable questions (2) ontologies enable standard reasoning and querying services over the knowledge, providing a framework for generating questions more easily.

Many efforts in the ontology-based QG are accompanied by methods for automating the task of difficulty-level estimation. In the E-ATG system [21], a state-of-the-art QG system, we have proposed an interesting method for predicting difficulty-level of the system generated factual questions. To recall, in that method, we assign a relatively high difficulty score to a question, if the concepts and roles in the question form a rare combination/pattern. For example, considering movie domain, if a question contains the roles: *is based on* and *won oscar*, which rarely appear together, the question is likely to be more difficult than those questions which are formed using a common role combination, say, *is directed by* and *is produced by*. Even though this method can correctly predict the difficulty-levels to a large extent, there are cases where this method fails. This is because there are other factors which influence the difficulty-level of a question.

An early effort to identify factors that could potentially predict the difficulty-level was by Seyler et. al [17,18]. They have introduced a method to classify a question as *easy* or *hard* by finding the features of the similar question entities in the Linked Open Data (LOD). Feature values for the classification task are obtained based on the connectivity of the question entities in the LOD. We observed that, rather than mapping to LOD – which is not always possible in the case of highly specific domains/domain-entities – incorporating domain knowledge in the form of terminological axioms and following an educational theory called Item Response Theory (IRT), the prediction can be made more accurate.

The contributions of this paper can be listed as follows.

- We reformulate some of the existing factors/features and propose new factors which influence the difficulty-level of a question, by taking into account the learners' knowledge level (or learners' category).
- We introduce ontology-based metrics for finding the feature values.
- With the help of standard feature selection methods in machine learning and by using a test dataset, we study the influence of these factors in predicting hardness of a question for three standard learner categories.
- We then propose three learner-specific regression models trained only with the respective influential features and, the output of the models is interpreted using the IRT to find the overall difficulty-level of a question.

This paper is organized as follows. Section 2 contains the preliminaries required for understanding the paper. Section 3 discusses the outline of the proposed method. In Section 4, we give an account of the related works. Section 5 proposes the set of features of a question which determines its difficulty-level. In Section 6, we explain the machine learning methods that we have adopted to develop the Difficulty-level Model (DLM). Further, we discuss the performance of DLM in Section 6.1. A comparison with the state-of-the-art method is given in Section 7. Conclusions and future line of research are detailed at the end.

2. Preliminaries

We assume the reader to be familiar with Description Logics[9] (DLs). DLs are decidable frag-

ments of first-order logic with the following building blocks: unary predicates (called *concepts*), binary predicates (called *roles*), instances of concepts (called *individuals*) and values in role assertions (called *literals*). A DL ontology is thought of as a body of knowledge describing some domain using a finite set of DL axioms. The concept assertions and role assertions form the assertion component (or ABox) of the ontology. The concept inclusion, concept equality, role hierarchy etc. (the type of axioms depend on the expressivity of the DL) form the terminological component (or TBox) of the ontology.

2.1. Question generation using patterns

For a detailed study of difficulty-level estimation, we use the *pattern-based* method, employed in the E-ATG system, for generating factual questions from the ABox of the given ontologies.

In the pattern-based question generation, a question can be considered as a set of *conditions* that asks for a solution which is explicitly present in the ontology. The set of conditions is formed using different combinations of concepts and roles assertions associated with an individual in the ontology. Example-1 is on such question, framed from the following assertions that are associated with the (*key*) individual `birdman`.

```
Movie(birdman)
isDirectedBy(birdman,alejandro)
hasReleaseDate(birdman,"Aug 27 2014")
```

Example 1 Name the *Movie* that *is directed by Alejandro* and *has release date Aug 27, 2014*.

For generating a question of the above type, we may need to use a (generic) SPARQL query template as shown below. The resultant tuples are then associated with a question pattern (E.g., Name the [?C], that is [?R1] [?o1] and [?R2] [?o2]. (key: ?s)) to frame the questions.

```
SELECT ?s ?C ? R1 ?o1 ?R2 ?o2   WHERE
{
    ?s a ?C .  ?s ?R1 ?o1 .  ?s ?R2 ?o2 .
    ?R1 a owl:ObjectProperty .
    ?R2 a owl:DatatypeProperty .
}
```

In [21], the authors have studied all the possible generic question patterns that are useful in generating common factual questions. They have also proposed methods for selecting *domain-relevant* resultant tuple55s/questions for conducting domain related assessments. A resultant tuple of the above query (for example, `?s = birdman, ?C = Movie, ?R1 = isDirectedBy, ?o1 = alejandro, ?R2 = hasReleaseDate, ?o2 = "Aug 27 2014"`) can be represented in the form of a set of triples ({`(birdman, a, Movie), (birdman, isDirectedBy, alejandro), (birdman, hasReleaseDate, "Aug 27 2014")`}). These triples, without the key, give rise to concept expressions that represent the conditions in the question. For example, the concept expression of "`(___, a, Movie)`" is the concept `Movie` itself. Similarly, the concept expression of "`(___, isDirectedBy, alejandro)`" is ∃`isdirectedBy.{alejandro}`. The conditions for the question given in Example-1 are:

Conditions: `Movie`, ∃`isdirectedBy.{alejandro}`, ∃`hasReleaseDate.{"Aug 27 2014"}`

It should be noted that, ∃`directedBy.{alejandro}` does not imply that the movie is directed *only* by Alejandro, but it is mandatory that he should be a director of the movie.

For the ease of understanding, all examples presented in this paper are from the Movie domain.

2.2. Item Response Theory

Item Response Theory (IRT) [14] models relationship between the ability or trait of a person and his responses to the *items* in an experiment. The term *item* denotes an entry, statement or a question used in the experiment. The item response can be *dichotomous* (yes or no; correct or incorrect; true or false) or *polytomous* (more than two options such as rating of a product). The quality measured by the item may be knowledge proficiency, aptitude, belief or even attitude. This theory was first proposed in the field of psychometrics, later, the theory was employed widely in educational research to calibrate and evaluate questions items in the world-wide examinations such as the Scholastic Aptitude Test (SAT) and Graduate Record Examination (GRE) [8].

In our experiments, we use the simplest IRT model often called *Rasch model* or the *one-parameter logistic model* (1PL) [11]. According to this model, a learner's response to a question item is determined by her knowledge proficiency level (a.k.a. *trait level*) and the difficulty of the item. 1PL is expressed in terms of the probability that

a learner with a particular trait level will correctly answer a question that has a particular difficulty-level. [14] represents this model as:

$$P(R_{li} = 1|\theta_l, \alpha_i) = \frac{e^{(\theta_l - \alpha_i)}}{1 + e^{(\theta_l - \alpha_i)}} \quad (1)$$

In the equation, R_{li} refers to the response (R) made by the learner l for the question item i (where $R_{li} = 1$ refers to a correct response), θ_l denotes the trait level of the learner l, α_i represents the difficulty score of item i. θ_l and α_i values are normalized to be in the range [-1.5 to 1.5]. $P(R_{li} = 1|\theta_l, \alpha_i)$ denotes the conditional probability that a learner l will respond to item i correctly. For example, the probability that a below-average trait level (say, $\theta_l = -1.4$) learner will correctly answer a question that has a relatively high hardness (say, $\alpha = 1.3$) is:

$$P = \frac{e^{(-1.4-1.3)}}{1 + e^{(-1.4-1.3)}} = \frac{e^{(-2.7)}}{1 + e^{(-2.7)}} = 0.063$$

In the paper, we intend to find the α_i of the factual questions which are meant for learners, whose trait levels are known to be either high, medium or low. We find the trait levels of the learners by gathering (and normalizing) their grades or marks obtained for a standard test of subject matter conducted in their enrolled institutions. The corresponding P values are obtained by finding the ratio of the number of learners (in the trait level under consideration) who have correctly answered the item, to the total number of learners at that trait level. On getting the values for θ_l and P, the value for α_i was calculated using the Equation-2.

$$\alpha_i = \theta_l - log_e(\frac{P}{1 - P}) \quad (2)$$

In the equation, $\alpha_i = \theta_l$, when P is 0.50. That is, a question's difficulty is defined as the trait level required for a learner to have 50 percent probability of answering the question item correctly. Therefore, for a trait level of $\theta_l = 1.5$, if $\alpha_i \approx 1.5$, we can consider that the question as having a high difficulty-level. Similarly, for a trait level of $\theta_l = 0$, if $\alpha_i \approx 0$, the question has a medium difficulty-level. In the same sense, for a trait level of $\theta_l = -1.5$, if $\alpha_i \approx -1.5$, then question has a low difficulty-level.

3. Outline of the proposed method

In this paper, based on the insights obtained by the study of the questions that are generated from the ATG[20] and E-ATG systems, we propose features/factors that can positively or negatively influence the difficulty-level of a question. Albeit there are existing methods which utilize some of these factors for predicting difficulty-level, studying the psychometric aspects of these factors by considering learners' perspective about the question, has given us further insight into the problem.

As we saw in Section 2.2, IRT is an item oriented theory which could be used to find the difficulty-level of a question by knowing the question's hardness (difficult or not difficult) with respect to various learner categories. Therefore, on finding the hardness of a given question based each on learner category, we can effectively use the IRT model for interpreting its overall difficulty-level.

According to IRT, a question is assigned a *high* difficulty-level if it is difficult for an expert learner to answer it correctly. A question is said to be difficult for an expert if the probability of a group of expert learners answering the question correctly is ≤ 0.5. Similarly, a question can be assigned a *medium* and *low* difficulty-level if the probability with which the question is answered by a group of intermediate learners is ≤ 0.5 and a group of beginner level learners is ≤ 0.5, respectively. Table 1 shows the difficulty-level assignment of three questions: Q_1, Q_2 and Q_3, based on whether they are difficult (denoted as d) or not difficult (represented as nd) for three learner categories.

Table 1
Assigning one of the three difficulty-levels: *high, medium* and *low*, by considering whether the question is difficult (d) or not-difficult (nd) for three learner categories.

Qn.	Expert	Intermed.	Beginner	Difficulty -level
Q_1	d	d	d	*high*
Q_2	nd	d	d	*medium*
Q_3	nd	nd	d	*low*

We consider three standard categories of learners: *beginners, intermediates* and *experts*, and model three classifiers for predicting the difficulty corresponding to the three learner categories, as shown in Fig. 1. Since the hardness (d/nd) corre-

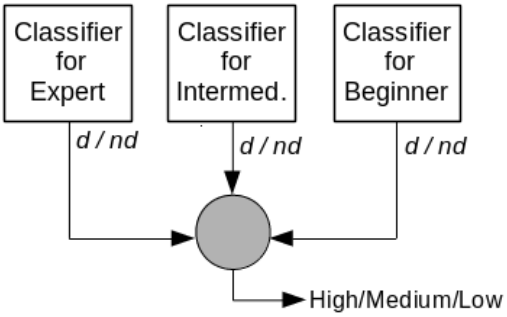

Fig. 1. Block diagram of the proposed model for predicting a question's difficulty-level

sponding to the three categories of learners should be predicted first from the feature values, machine learning models/classifiers which can learn from available training data is an obvious choice. We consider only those factors which are influential for a given learner category for training the models. The output of the three classifiers is matched with the content of Table 1 to find the question's overall difficulty-level.

4. Related Work: Difficulty-level Estimation

A simple notion to find the difficulty-level of an ontology-generated multiple choice questions (MCQs) was first introduced by Cubric and Tosic[10]. Later, in [3], Alsubait et al. extended the idea and proposed a similarity-based theory for controlling the difficulty of ontology-generated MCQs. In [6], they have applied the theory on analogy type MCQs. In [4], the authors have experimentally verified their approach in a student-course setup. The practical solution which they have suggested to find out the difficulty-level of an MCQ is with respect to the degree of similarity of the distractors to the key. If the distractors are very similar to the key, students may find it very difficult to answer the question, and hence it can be concluded that the MCQ is difficult.

In many a case, the question statement in an MCQ is also a deciding factor for the difficulty of an MCQ. For instance, the predicate combination or the concepts used in a question can be chosen such that they can make the MCQ difficult or easy to answer. This is the reason why in this paper we focus on finding difficulty-level of questions having no choices (i.e., non-MCQs). An initial investigation of this aspect was done in [21]. Concurrently, there was another relevant work by Seyler et. al[17,18], focusing on QG from knowledge graphs (KGs) such as DBpedia. For judging the difficulty-level of such questions, they have designed a classifier trained on Jeopardy! data. The classifier features were based on statistics computed from the KGs (Linked Open Data) and Wikipedia. However, they have not considered the learner's knowledge level, as followed in the IRT, while formulating the feature metrics. This makes their measures less employable in sensitive applications such as in an e-Learning system. While considering ontology-based questions, one of the main limitation of their approach is that the feature values were determined based on the connectivity of question entities in the KG, whereas in the context of DL ontologies, the terminological axioms can be also incorporated to derive more meaningful feature metrics. In addition, the influence of the proposed factors in determining the difficulty using feature selection methods was not studied.

5. Proposed Factors to determine Difficulty-level of Questions

In this section, we look at a set of factors which can possibly influence the difficulty-level of a question and propose ontology-based metrics to calculate them. The intuitions for choosing those factors are also detailed.

To recall, a given question can be thought of as a set of conditions. For example, consider the following questions (where the underlined portions denote the equivalent ontology concepts/roles used).

Qn-1: *Name the Movie that was directed by Clint Eastwood.*

Qn-2: *Name the Oscar movie that was directed by Clint Eastwood.*

The equivalent set of conditions of the two questions can be written as:

Conditions in Qn-1: `Movie,`
$\exists$`directedBy.{clint_eastwood}`

Conditions in Qn-2: `Oscar_movie,`
$\exists$`directedBy.{clint_eastwood}`

5.1. Popularity

Popularity is considered as a factor because of the intuition that the greater the popularity of

the entities that form the question, more likely that a learner answers the question correctly. (We observe that this notion is applicable for learners of all categories.) Therefore, the question becomes easier to answer if the popularity of the concepts and roles that are present in the question is high. For example, out of the following two questions, Qn-3 is likely to be easy to answer than Qn-4, since `Oscar_movie` is a popular concept than `Thriller_movie`.

Qn-3: *Name an oscar movie.*
Qn-4: *Name a thriller movie.*

Our approach for measuring popularity is based on the observation that, (similar to what we see in Wikipedia data) if more articles talk about a certain entity, the more important, or popular, this entity is. In Wikipedia, when an article mentions a related entity, it is usually denoted by a link to the corresponding Wikipedia page. These links form a graph which is exploited for measuring the importance of an entity within Wikipedia. Keeping this in mind, we can define the popularity of an entity (individual) in an ontology as the number of object properties which are linked to it from other individuals. For obtaining a measure in the interval [0,1], we divide the number of in-links by the total amount of individuals in the ontology.

To find the popularity of a concept C in ontology $\mathcal{O}$, we find the mean of the popularities of all the individuals which satisfy C in $\mathcal{O}$. If the condition in a question is a role restriction, then the concept expression of it will be considered, and popularity is calculated. The overall popularity of the question is determined by taking the mean of the popularities of all the concepts and role restrictions present in it.

5.2. Selectivity

Selectivity of the conditions in a question helps in measuring the quality of the hints that are present in it [17]. Given a condition, selectivity refers to the number of individuals that satisfy it. When the selectivity is high, a question tends to be easy to answer. For example, among the following questions, clearly, Qn-5 is easier to answer than Qn-6. This is because finding an actor who has acted in at least a movie is easy to answer than finding an actor who has acted in a particular movie; finding the latter requires more specific knowledge.

Qn-5: *Name an actor who acted in a movie.*
Qn-6: *Name an actor who acted in Argo.*

To formalize such a notion, we can look at the *answer space* corresponding to each of the conditions in the questions. Answer space simply denotes the *count of individuals* satisfying a given condition. We will represent answer space of a condition c as $ASpace(c)$.

The conditions in the above questions are:

Conditions in Qn-5: `Actor`, $\exists$`actedIn.Movie`
Conditions in Qn-6: `Actor`, $\exists$`actedIn.{argo}`

Since $ASpace(\exists$`actedIn.{argo}`$)$ is very much lesser than $ASpace(\exists$`actedIn.Movie`$)$, we can say that Qn-6 is difficult to answer than Qn-5. (Actors who acted only in dramas are not possible answers to Qn-5.)

As a question can have more than one conditions present in it, answer spaces of all the condition have to be taken into account while calculating the overall difficulty score of the question. It is debatable that including a specific condition in the question can always make the question difficult to answer – sometimes a specific condition can give a better hint to a (proficient) learner.

For example, the following question is more difficult to answer than Qn-5 and Qn-6 for a non-expert, since $ASpace($`American_actor`$) << ASpace($`Actor`$)$.

Qn-7: *Name an American actor who acted in Argo.*

However, for an expert, given that the actor is an American is an additional hint, making the question sometimes easier than Qn-5 and 6. Therefore, we can roughly assume the relation between difficulty-level and answer space as follows, where D_{expert} and $D_{beginner}$ correspond to the difficulty-level for an expert learner and difficulty-level for a beginner respectively. We will closely look at these relations in the following subsections.

$$D_{expert} \propto ASpace$$

$$D_{beginner} \propto \frac{1}{ASpace}$$

When a question contains multiple conditions, we do an aggregation of their normalized (or relative) answer spaces (denoted as $RASpace$) to find the overall answer space (addressed as $ASpaceOverall$) of the question. We find the $RASpace$ of a concept by dividing the count of in-

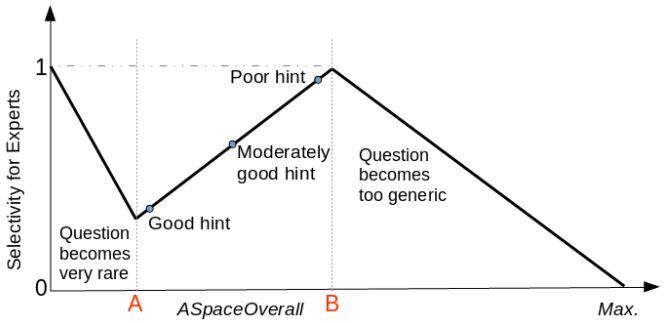

Fig. 2. Relation between selectivity and answer space for experts

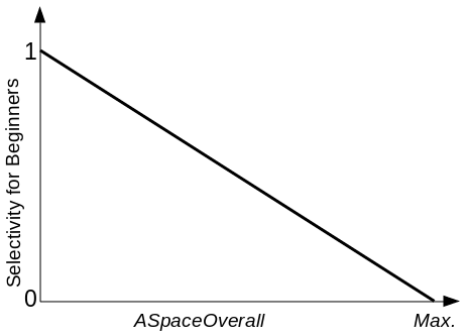

Fig. 3. Relation between selectivity and answer space for beginners

dividuals satisfying the concept by the total count of individuals in the apex concept (Thing class) of the ontology. For instance, $RASpace(\{\texttt{argo}\}) = ASpace(\{\texttt{argo}\})/ASpace(\texttt{owl:Thing})$. Similarly, if the condition is a role related restriction, corresponding domain concept of the role will be used to find the relative answer space. For $\exists\texttt{actedIn.}\{\texttt{argo}\}$, *RASpace* is calculated as: $ASpace(\exists\texttt{actedIn.}\{\texttt{argo}\})/ASpace(\text{Domain}(\texttt{actedIn}))$. The overall answer space can be found by taking the average of all the relative answer spaces of the conditions in the question, where $C_S = \{t_1, t_2, ..., t_n\}$ is the set of conditions in the question S, and $|C_S| = n$.

$$ASpaceOverall(C_S) = \frac{\sum_{i=1}^{n} RASpace(t_i)}{n} \quad (3)$$

In the following paragraphs, we discuss how the selectivity feature would affect the difficulty-level of an item. We discuss the cases of expert, intermediate and beginner learners separately. In the process, we define two selectivity based features and specify how to compute them using the knowledge base and the domain ontology.

Expert learner An expert learner is assumed to have a well developed structured knowledge about the domain of discourse. She is supposed to clearly distinguish the terminologies of the domain and is capable of doing reasoning over them. Therefore, in general, selectivity can be assumed to be directly proportional to the difficulty-level; that is, when the *ASpaceOverall* increases, the underlying hints becomes poor and the question is likely to become difficult for her. However, intuitively, below and beyond particular *ASpaceOverall* values, a question's difficulty does not necessarily follow this proportionality. As pointed out in [21,20] when a question pattern becomes rare, it becomes difficult to answer the question correctly. Therefore, in Fig. 2, towards the left of the point A, the question tends to become difficult, since the answer space becomes too small. Similarly, towards the right of the point B, the question tends to become more generic and its difficulty diminishes. To accurately predict whether a question is difficult or not, it is necessary to statistically determine the positions of the points A and B. Based on the initial analysis of the empirical data obtained from [21], we processed with an assumption that the question tends to become too generic when the *ASpaceOverall* $\geq 50\%$ of the total number of individuals in the ontology. Similarly, the question starts to become difficult when the *ASpaceOverall* $\leq 10\%$ of the total number of individuals. The selectivity corresponding to an expert is expressed as $Selectivity_{Ex}$. Knowing the overall answer space of a question, selectivity is computed directly from the graph in Fig. 2 – in the graph, Max, A(10%) and B(50%) are known points.

Beginner learner A beginner is assumed to have a less developed internal knowledge structure. She can be assumed to be familiar with the generic (sometimes popular) information about the domain and is less aware about the detailed specifics. We assume that the *selectivity* factor behaves proportionally to the *ASpaceOverall*, unlike what we saw in the experts' case. The intuition behind this assumption is that, when the overall answer space increases, as in the case of an expert the so-called hints in the question cannot be expected to become poor; this is because, a person with poorly developed domain knowledge may not be able to

differentiate the quality or property of the hint, making it rather a factor for generalizing the question (thereby making the question easily answerable). Therefore, we can follow a linear proportionality relation as shown in Fig. 3, to find the difficulty for a beginner, and we can denote this new selectivity as $Selectivity_{Bg}$.

Intermediate learner An intermediate learner can be assumed to have partially both the perspective of an expect as well that of a beginner. Therefore, we can assume her selectivity value as combination of $Selectivity_{Ex}$ and $Selectivity_{Bg}$ – considering them as two factors.

5.3. Coherence

In the current context, coherence captures the semantic relatedness of entities (individuals and concepts) in a question. It can be best compared to measuring the co-occurrences of individuals and concepts in the text. While considering coherence as a factor, we assume that higher the coherence between individuals/concepts in a question, lower is its difficulty-level and vice versa, because intuitively, the facts about highly coherent entities are likely to be recalled easier than the facts about less coherent entities. It is observed that this notion is applicable for learners of all categories.

Qn-8: *Name the hollywood-movie starring Anil Kapoor and Tom Cruise.*
Qn-9: *Name the hollywood-movie starring Tom Cruise and Tim Robbins.*

Considering the above two questions, coherence between the concept `HollywoodMovie` and the individuals: `anil_kapoor, tom_cruise`, is lesser (since there is only one movie they both have acted together) than the coherence between `HollywoodMovie, tom_cruise` and `tim_robbins`, making the former question difficult to answer than the latter.

Given an ontology, we measure the coherence between two of its individuals as the sum of the ratio between the size of the set of entities that point to both individuals and the size of the union of the sets of entities that point to either one of the individuals, and the ratio between the size of the set of entities that are pointed by both individuals and the size of the union of the sets of entities that are pointed by either one of the individuals. Formally, the coherence between two individuals p and q can be represented as in Eq. 4, where I_i is the set of entities from which the individual i is having incoming relations and O_i is the set of entities to which i is having outgoing relations.

$$Coherence(p,q) = \frac{|I_p \cap I_q|}{|I_p \cup I_q|} + \frac{|O_p \cap O_q|}{|O_p \cup O_q|} \quad (4)$$

Each portion of the measure is known as the Jaccard similarity coefficient, which is a statistical method to compare the similarity of sets.

5.4. Specificity

Specificity refers to how specific a question is. For example, among the following questions, Qn-2 is more specific question than Qn-10 and requires more knowledge proficiency to answer it correctly. We consider Qn-2 as more difficult to answer than Qn-10.

Qn-2: *Name an Oscar movie that was directed by Clint Eastwood.*
Qn-10: *Name the movie that is related to Clint Eastwood.*

For a learner, the difficulty-level depends on how detailed the question is. Intuitively, if a question contains domain specific conditions, the probability of a learner for correctly answering the question will reduce. (This notion is observed to be applicable for all categories of learners.) To capture this notion, we utilize the concept and role hierarchies in the domain ontology. We relate the depths of the concepts and roles that are used in the question to the concept and role hierarchies of the ontology, to determine the question difficulty. To achieve this, we introduce *depthRatio* for each predicate p in an ontology. *depthRatio* is defined as:

$$depthRatio_{\mathcal{O}}(p) = \frac{\begin{array}{c}\text{Depth (or length) of } p \\ \text{from the root of the hierarchy}\end{array}}{\begin{array}{c}\text{Maximum length of} \\ \text{the path containing } p\end{array}} \quad (5)$$

For a question S, generated from an ontology $\mathcal{O}$, with x as key and P as the set of concepts/roles in S, let $\mathcal{C}$ denote the set of concepts satisfied by x, and let $\mathcal{R}$ represents the set of roles such that either x is present at their domain (subject) or range (object) position (i.e., $R \in \mathcal{R} \implies \mathcal{O} \models$

$R(x,i) \vee R(i,x)$, where i is an arbitrary instance in $\mathcal{O}$). For each $p \in P$, we find the largest subset in $\mathcal{C}$ (if p is a concept) or we find the largest subset in $\mathcal{R}$ (if p is a role), such that the elements in the subset can be related using the relation $\sqsubseteq$, and p is an element in that subset. The cardinality of such a subset forms the denominator of Eq. 5, and the numerator is the position of the predicate p from the right (right represents the top concept or top role) when the elements in the subset are arranged using the relation $\sqsubseteq$.

A stem can have more than one predicate present in it. In that case, we assume that the predicate with a highest depthRatio (associated with the reference individual) could potentially make the stem more specific. Therefore, we define the overall depthRatio of a stem (called the *specificity*) as the product of the average depthRatio with the maximum of all the depthRatios.

6. Difficulty-level Modeling of Questions

In the previous section, we have proposed a set of features which possibly influence the difficulty-level of a question. In this section, we do a feature selection study using three widely used filter models to find out the amount of influence of the proposed factors in predicting question difficulty. We then train three logistic regression models (RM_e, RM_i, RM_b) for each learner category (experts, intermediates and beginners, respectively) using the selected prominent features. Their predictions for a given question are taken to find the overall difficulty-level. Ten-fold cross validation is used to find the performance of the three models.

Training data The training data consisted of a set of 520 questions that were generated from four ontologies (DSA, MAHA, GEO and PD ontologies – see our project website[4] for details) available online. These questions were classified as *difficult* or *not-difficult* for each of the three learner categories (we denote the training data for experts, intermediate and beginners respectively as TD_e, TD_i and TD_b). The classification is done by either of the two ways, (1) in a classroom setting by using IRT or (2) with the help of subject matter experts. In the former case, we find the probability by which a particular question is answered correctly by a learner of specific knowledge proficiency level and assign it as difficulty (d) or not (nd). In the latter case, more than 5 domain experts were asked to do the ratings and their majority ratings were considered for assigning d or nd. All the question that were used for training had been previously used as benchmark sets in [12,21,20]. In the training data, the question identifiers are accompanied by five feature values tabulated from the respective ontologies along with their difficulty assignment. The feature values are normalized to values between 0 and 1. An instance of the training data is given in Fig. 4.

```
Item identifier: dsa_1
Popularity: 0.231
Selectivity_Ex: 0.320
Selectivity_Bg: 0.113
Coherence: 0.520
Specificity: 0.440
Difficulty: d
```

Fig. 4. An instance of the training data

Feature Selection In order to find out the amount of influence of each of the proposed factors, we did an attribute evaluation study using three popular feature selection approaches: Information Gain[15] (IG), ReliefF[16] (RF) and Correlation-based[13] (CB) methods. These feature selection approaches select a subset of features that minimize redundancy and maximize relevance to the target such as the class labels in classification. The ranking scores/weights obtained for the features are given in Table 2.

In Table 2, we can see that, the least prominent feature for finding the difficulty for experts is the Selectivity$_{Bg}$, since all the three filter models ranked it as the least influential one – see the fields shaded in blue in the three TD_e columns. In the case of predicting difficulty for intermediates, the ranking scores of Selectivity$_{Ex}$ is less than that of Selectivity$_{Bg}$ when the models used are RF and CB – see the fields shaded in red. When it comes to beginner learners, the factor Selectivity$_{Ex}$ is found to have the least influence – see the fields shaded in gray. While developing the IDM, we have ignored the least influential features for training the regression models.

[4] Project website: https://sites.google.com/site/ontoassess/

Table 2
Ranking score of features for the three training sets using three popular filter models. (IG, RF and CB, denote the three filter models: Information Gain, ReliefF and Correlation-based, respectively.)

	IG			RF			CB		
	TD_e	TD_i	TD_b	TD_e	TD_i	TD_b	TD_e	TD_i	TD_b
Popularity	0.7613	0.6344	0.7925	0.831	0.378	0.178	0.766	0.621	0.562
Selectivity$_{Ex}$	0.8802	0.6913	0.0963	0.738	0.451	0.120	0.699	0.258	0.058
Selectivity$_{Bg}$	0.0012	0.6553	0.9996	0.008	0.668	0.177	0.114	0.442	0.249
Coherence	0.5638	0.3251	0.8112	0.761	0.487	0.211	0.731	0.538	0.315
Specificity	0.6577	0.5436	0.5751	0.651	0.521	0.459	0.657	0.761	0.602

Observations Consistent to what we have postulated in Section 5.2, Selectivity$_{Ex}$ is found to be a more influential factor than Selectivity$_{Bg}$, for deciding the difficulty of a question for an expert learner. Similarly, for a beginner, Selectivity$_{Bg}$ is found to be more influential than Selectivity$_{Ex}$.

6.1. Performance of regression models

The performances of three learner-specific regression models: RM_e, RM_i, RM_b, considering all the 5 features are 76.73%, 78.6% and 84.23% respectively. These percentage values indicate the ratio of the number of instances classified correctly to the total number of instances given for classification, under 10-fold cross-validation setting. After removing the least influential features, the performance of the classifiers became 76.9%, 79.8%, and 85% respectively. The difference in the performance before and after feature selection is roughly the same because the model can theoretically assign minimum or zero weight to non-influential features. However, we did the feature selection and ranking to evaluate our hypothesis about what features are influential in which case.

When the overall system was run on the available dataset of factual questions from different domains (520 questions), it is observed that DLM correctly classifies about 77% of them. These questions are relevant to the domain selected using the heuristics given in [21].

6.2. Non-classifiable Questions

Following from what we have seen in Section 3, the DLM could not assign a difficulty-level to a given question if the outcomes of the three regression models do not agree with the three possible assignments (see Fig. 1). We call such questions as *non-classifiable* ones and the others as *classifiable* questions. We investigated the percentage of such non-classifiable cases by analyzing questions generated from five ontologies available online (available in our project website). We used questions that were generated in [20] for our study, the statistics of the non-classifiable cases are given in Fig. 5 (Note that the X-axis begins with 50%). On an average, 10% of all the questions generated from an ontology are found to be non-classifiable.

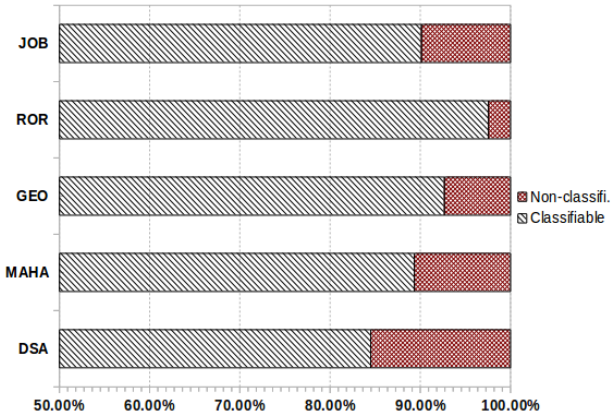

Fig. 5. Classifiable Vs Non-classifiable questions

An analysis of the non-classifiable questions taken from the DSA ontology shows that incompleteness of the ontology and the way the domain is modeled as an ontology are the main reasons for this discrepancy. For example, *Name a doubly linked list* is a question which is assigned to be difficult for an expert (since the Doubly Linked List concept has only one individual and the incompleteness of data makes it a less popular concept), and not difficult for a beginner (the depthRatio is high for the concept `Doubly_Linked_List` because, it forms a specific concept in the ontology. However, this issue would not have appeared, if

the concept `Doubly_Linked_List` had been modeled as an individual.

7. Comparison with existing method

In this section, we compare the predictions of difficult-levels by the proposed model and the method given in [21]. We call the latter as *E-ATG method*. We do not report a comparison with the model proposed in [17,18] because their difficulty-level model is not a domain ontology-based model and prediction is possible only if the question components can be mapped to Linked Open Data entities. In addition, they could predict the question difficulty either as *easy* or *hard*, whereas our model classifies the question into three standard difficulty-levels: *high, medium* and *low*.

In [21], effectiveness of E-ATG method is established by comparing the predicated difficulty-levels with their actual difficulty-levels determined in a classroom setting. DSA ontology was used for the study. Twenty four representative questions (given in Appendix A), selected from 128213 generated questions, were utilized for the comparison. (More details about the selection process can be found at [21]). A correlation of 67% between the predicted and actual difficulty-levels was observed[5]. We tested the approach proposed in this paper on the above set of questions.

While comparing the proposed difficulty-levels with the actual difficult-levels, we found that 21 out of 24 are matching (87.5% correlation), and one benchmark question is identified as non-classifiable.

7.1. Discussion

The E-ATG method mainly considered only one feature, the *triviality score* (which denotes how rare the property combination in the stem are), for doing the predication. Our results (8.5% improvement) show that the proposed set of new features could improve the correctness of the prediction. The current model is trained only using 520 training samples. We expect the system to perform even better after training with more data as and when they are available, and by identifying other implicit features. Due to unavailability of large training data, unsupervised feature learning methods cannot be effectively applied in this context.

[5]To get more accurate result, the calculations were redone with θ values between: [-1.5, 1.5], and 1.25, 0 and -1.25 as the medians of the α values for experts, beginners and intermediates, respectively with $\pm.25$ standard deviation

8. Conclusions and Future Work

Establishing mechanisms to control and predict the difficulty of assessment questions is clearly a big gap in existing question generation literature. Our contributions have covered the deeper aspects of the problem, and proposed strategies, that exploit ontologies and associated measures, to provide a better difficulty-level predicting model, that can address this gap. We developed the difficulty-level model (DLM) by introducing three learner-specific logistic regression models for predicting the difficulty of a given question for three categories of learners. The output of these three models was then interpreted using the Item Response Theory to assign *high, medium* or *low* difficulty-level. The overall performance of the DLM and the individual performance of the three regression models based on cross-validation were reported and they are found to be satisfactory. Comparison with the state-of-the-art method shows an improvement of 8.5% in correctly predicating the difficulty-levels of benchmark questions.

The model proposed in this paper for predicting the difficulty-level of questions is limited to ABox-based factual questions. It would be interesting to extend this model to questions that are generated using the TBox-based approaches. However, the challenges to be addressed would be much more, since, in the TBox-based methods, we have to deal with many complex restriction types (unlike in the case of ABox-based methods) and their influence on the difficulty-level of the question framed out of them needs a detailed investigation.

For establishing the propositions and techniques stated in this paper, we have implemented a system which demonstrates the feasibility of the methods on medium sized ontologies. It would be interesting to investigate the working of the system on large ontologies.

Acknowledgements

This project is funded by Ministry of Human Resource Development, Gov. of India. We express our fullest gratitude to the participants of our evaluation process: Dr. S.Gnanasambadan (Director of Plant Protection, Quarantine & Storage), Ministry of Agriculture, Gov. of India; Mr. J. Delince and Mr. J. M. Samraj, Department of Social Sciences AC & RI, Killikulam, Tamil Nadu, India; Ms. Deepthi.S (Deputy Manager), Vegetable and Fruit Promotion Council Keralam (VFPCK), Kerala, India; Dr. K.Sreekumar (Professor) and students, College of Agriculture, Vellayani, Trivandrum, Kerala, India. We also thank all the undergraduate and post-graduate students of Indian Institute of Technology, Madras, who have participated in the empirical study.

References

[1] Asma Ben Abacha, Marcos Da Silveira, and Cédric Pruski. Medical ontology validation through question answering. In *AIME*, pages 196–205, 2013. 10.1007/978-3-642-38326-7_30.

[2] Maha Al-Yahya. Ontology-based multiple choice question generation. *The Scientific World Journal*, Vol 2014, page 9, ID: 10.1155/2014/274949, 2014.

[3] T. Alsubait, B. Parsia, and U. Sattler. A similarity-based theory of controlling mcq difficulty. In *e-Learning and e-Technologies in Education (ICEEE), 2013 Second International Conference on*, pages 283–288, Sept 2013.

[4] T. Alsubait, B. Parsia, and U. Sattler. Generating multiple choice questions from ontologies: Lessons learnt. In *Proceedings of the 11th International Workshop on OWL: Experiences and Directions (OWLED 2014)*, volume 1265, pages 73–84, Oct 2014.

[5] Tahani Alsubait. *Ontology-based multiple-choice question generation*. PhD thesis, School of Computer Science, The University of Manchester, 2015.

[6] Tahani Alsubait, Bijan Parsia, and Ulrike Sattler. Mining ontologies for analogy questions: A similarity-based approach. volume 849 of *CEUR Workshop Proceedings*. OWL Experiences and Directions, 2012.

[7] Tahani Alsubait, Bijan Parsia, and Ulrike Sattler. Generating multiple choice questions from ontologies: Lessons learnt. volume 1265 of *CEUR Workshop Proceedings*. OWL Experiences and Directions, 2014.

[8] Xinming An and Yiu-Fai Yung. Item response theory: What it is and how you can use the irt procedure to apply it. In *SAS Global Forum*, 2014.

[9] Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors. *The description logic handbook: theory, implementation, and applications*. Cambridge University Press, New York, NY, USA, 2003.

[10] Marija Cubric and Milorad Tosic. Towards automatic generation of e-assessment using semantic web technologies. In *Proceedings of the 2010 International Computer Assisted Assessment Conference*, 2010.

[11] Christine DeMars. *Item Response Theory: Understanding Statistics*. Oxford University Press, 2010.

[12] Vinu E.V. and Kumar P. Sreenivasa. A novel approach to generate mcqs from domain ontology: Considering dl semantics and open-world assumption. *Web Semantics: Science, Services and Agents on the World Wide Web*, 34:40 – 54, 2015. http://dx.doi.org/10.1016/j.websem.2015.05.005.

[13] Mark A Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, Department of Computer Science, The University of Waikato, 1999.

[14] R. Michael Furr and Verne R. Bacharach. *Psychometrics, An Introduction. Second Edition*. SAGE Publications, Inc, 2014.

[15] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1226–1238, August 2005.

[16] Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53(1):23–69, 2003.

[17] Dominic Seyler, Mohamed Yahya, and Klaus Berberich. Generating quiz questions from knowledge graphs. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 113–114, New York, NY, USA, 2015. ACM.

[18] Dominic Seyler, Mohamed Yahya, and Klaus Berberich. Knowledge questions from knowledge graphs. *CoRR*, abs/1610.09935, 2016.

[19] Dominic Seyler, Mohamed Yahya, Klaus Berberich, and Omar Alonso. Automated question generation for quality control in human computation tasks. In *Proceedings of the 8th ACM Conference on Web Science, WebSci 2016, Hannover, Germany, May 22-25, 2016*, pages 360–362, 2016.

[20] Ellampallil Venugopal Vinu and Puligundla Sreenivasa Kumar. Improving large-scale assessment tests by ontology based approach. In *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2015, Hollywood, Florida. May 18-20, 2015.*, page 457, 2015.

[21] E.V Vinu and P. Sreenivasa Kumar. Automated generation of assessment tests from domain ontologies. *Semantic Web Journal*, In Press, 2016.

[22] Branko Žitko, Slavomir Stankov, Marko Rosić, and Ani Grubišić. Dynamic test generation over ontology-based knowledge representation in authoring shell. *Expert Systems with Applications*, 36(4):8185 – 8196, 2009. http://dx.doi.org/10.1016/j.eswa.2008.10.028.

[23] Konstantinos Zoumpatianos, Andreas Papasalouros, and Konstantinos Kotis. Automated transformation of swrl rules into multiple-choice questions. In R. Charles Murray and Philip M. McCarthy, editors, *FLAIRS Conference*. AAAI Press, 2011.

Appendix A

Tables 3-5 contain benchmark stems that are generated from the Data Structures and Algorithms (DSA) ontology. In those tables, the stems 1 to 6, 7 to 16 and 17 to 24 correspond to high, medium and low (actual) difficulty-levels respectively. These stems were employed in the experiment mentioned in Section 7. Stems- 7, 8 and 20 are the uncorrelated ones. Stem-20 is identified as non-classifiable questions by our approach.

Table 3
Questions generated from the DSA ontology that are having *high* actual difficulty-levels.

Item No.	Stems of MCQs
1	Name a polynomial time problem with application in computing canonical form of the difference between bound matrices.
2.	Name an NP-complete problem with application in pattern matching and is related to frequent subtree mining problem.
3.	Name an all pair shortest path algorithm that is faster than Floyd-Warshall Algorithm.
4.	Name an application of an NP-complete problem that is also known as Rucksack problem.
5.	Name a string matching algorithm that is faster than Robin-Karp algorithm.
6.	Name a polynomial time problem that is also known as maximum capacity path problem.

Table 4
Questions generated from the DSA ontology that are having *medium* actual difficulty-levels.

Item No.	Stems of MCQs
7.	Name an NP-hard problem with application in logistics.
8.	Name the one whose worst time complexity is n exp 2 and with Avg time complexity n exp 2.
9.	Name the one which operates on output restricted dequeue and operates on input restricted dequeue.
10.	Name the operation of a queue that operates on a priority queue.
11.	Name a queue operation that operates on double ended queue and operates on a circular queue.
12.	Name the ADT that has handling process "LIFO".
13.	Name an internal sorting algorithm with worse time complexity m plus n.
14.	Name a minimum spanning tree algorithm with design technique greedy method.
15.	Name an internal sorting algorithm with time complexity n log n.
16.	Name an Internal Sorting Algorithm with worse time complexity n exp 2.

Table 5
Questions generated from the DSA ontology that are having *low* actual difficulty-levels.

Item No.	Stems of MCQs
17.	Name a file operation.
18.	Name a heap operation.
19.	Name a tree search algorithm.
20.	Name a queue with operation dequeue.
21.	Name a stack operation.
22.	Name a single shortest path algorithm.
23.	Name a matrix multiplication algorithm.
24.	Name an external sorting algorithm.