Automating the Generation of Competency Questions for Ontologies with AgOCQs

Mary-Jane Antia[1][0000−0002−9983−6267] and C. Maria Keet[1][0000−0002−8281−0853]

University of Cape Town, Cape Town, South Africa
{mantia,mkeet}@cs.uct.ac.za

Abstract. Competency Questions (CQs) are natural language questions drawn from a chosen subject domain and are intended for use in ontology engineering processes. Authoring good quality and answerable CQs has been shown to be difficult and time-consuming, due to, among others, manual authoring, relevance, answerability, and re-usability. As a result, few ontologies are accompanied by few CQs and their uptake among ontology developers remains low. We aim to address the challenges with manual CQ authoring through automating CQ generation. This novel process, called AgOCQs, leverages a combination of Natural Language Processing (NLP) techniques, corpus and transfer learning methods, and an existing controlled natural language for CQs. AgOCQs was applied to CQ generation from a corpus of Covid-19 research articles, and a selection of the generated questions was evaluated in a survey. 70% of the CQs were judged as being grammatically correct by at least 70% of the participants. For 12 of the 20 evaluated CQs, the ontology expert participants deemed the CQs to be answerable by an ontology at a range of 50%-85% across the CQs, with the remainder uncertain. This same group of ontology experts found the CQs to be relevant between 70%-93% across the 12 CQs. Finally, 73% of the users group and 69% of the ontology experts judged all the CQs to provide clear domain coverage. These findings are promising for the automation of CQs authoring, which should reduce development time for ontology developers.

1 Introduction

Ontologies have been shown to be useful in a wide range of subject domains and applications. Regarding their content, they are expected to be well-delineated, explicit, and representative of the selected subject domain. To obtain those characteristics, Competency Questions (CQs) have been proposed as important means, to aid in the development, verification, and evaluation processes [5, 13, 22, 25]. CQs are natural language questions drawn from a given (sub)domain for use in the ontology development cycle [22]. The adoption of CQs by ontology engineers has been reported as low due to difficulties including the authoring of good quality CQs [12]. In solving problems related to CQs for the ontology development cycle, the focus of several CQ-related studies has been on artefacts and processes that can enhance CQ quality after they have been manually authored [3, 12, 19, 22] Corpus-based methods have been used in several areas such

as expert systems and data mining [16, 24]. They are known to provide insights into knowledge domains, yet they have not been used as for CQ development. We aim to reduce hurdles with manual CQ authoring and quality by proposing an approach that uses corpus-based methods to automate CQ authoring. We aim to answer the following research questions:

Q1: Can a corpus-based method support the automation of CQ authoring, and if so, how?

Q2: How do automated CQs fare on the key quality criteria answerability, grammaticality, scope, and relevance in relation to a given (sub)domain?

The approach taken to answer these questions is as follows. We design a novel pipeline and accompanying algorithm to automate CQ generation. It combines a text corpus with CQ templates, a CQ abstraction method, transfer learning models, and NLP techniques. This procedure was evaluated with a survey among the target groups composed of ontology experts, domain experts, and users. The results showed that 14/20 CQs had 70%-100% of participants judged it as grammatically correct. Perceived answerability by a hypothetical ontology gave mixed results. Also, the vast majority of participants found 12/20 CQs to be relevant, and 73% of users and 69% of the experienced ontology experts found the CQs to provide clear domain coverage.

The rest of the paper discusses related work (Section 2), the novel methodology of AgOCQs (Section 3) and its evaluation (Section 4). We close with conclusions and future work (Section 5).

2 Related Work

We consider both the state-of-the-art CQs in ontology engineering and the promising NLP techniques for automated question generation.

2.1 CQs in Ontology Engineering

CQs have been proposed as part of the requirement specification in ontology development [23]. CQs have been shown to play other roles, such as functional requirements for ontology development, for verification (with a focus on completeness and correctness), and providing insights into the contents of a specific ontology, especially to non-expert users [5, 23]. CQs have also been used for ontology reuse and relevance [4], test-driven ontology development [13], and enhancing the agile development process for ontologies [1, 18].

A number of concerns have been noted in the literature around CQs in ontology development and use. In particular, manual authoring of CQs continues to be a hindrance to their successful participation in the ontology engineering process [3, 12, 19, 22, 26]. Many ontologies tend to have accompanying CQs defined at a high level to provide an overview of what the ontology is on [8]. There is a dearth of authoring support tools and the manual process of authoring CQs is tedious and time-consuming. In addition, manually authored CQs are sometimes not answerable, not relevant, not grammatical, and not sufficiently indicative

of the scope of a given ontology [3]. In a bid to address the lack of authoring support, several solutions have been proposed. Current solutions and tools have centred on evaluating manually CQs from specific ontologies [6, 5], creating artefacts to support manually developed CQs for ontology use [3, 12, 19, 22, 26]; or on how to develop ontologies from CQs [1]. One such solution is exemplified by a set of core and variant CQ archetypes created for the Pizza ontology [22]. These archetypes are ontology elements of OWL class and object property with a 1:1 mapping attribute, limiting their use to only OWL ontologies with certain limited formalisation patterns.

An approach is to check manually authored CQs with a few CQ patterns, focusing on OWL ontology variables [6] however, omit "Who" and "Where" question types. Another proposal separated the CQ linguistic analysis from OWL, using linguistic pattern extraction from CQs by [19, 26]. They identified nouns and noun phrases from entities along with verbs and verb phrases from predicates representing them in abstract forms for a set of 234 manual CQs that were developed and corrected for 5 different ontologies (Dem@care, Stuff, African wildlife, OntoDT and Software ontologies).

They also created 106 patterns [19, 26], from which the Competency question Language for specifying Requirements for an Ontology (CLaRO) controlled natural language was developed [12].

These templates restrict the CQs types but allow for a 1:m mapping and they can be used also with other ontology languages. CLaRO templates have shown good coverage with an accuracy of over 90% to unseen CQs across several domains [3, 12]. Each CLaRO template can correspond to several questions and are about 150 templates. However, for ontology developers to make use of them, CQs would still have to first be manually developed and then checked for compliance using the templates. The persistent manual-only approach is the common theme, and shortcoming, in the different solutions thus far.

2.2 Transfer Learning

With the advent of machine learning technologies, the process of question generation has had successes and failures. Transfer Learning (TL), a method which leverages knowledge learned in one domain task to perform a similar task in a different domain, has been at the centre of it [2, 10]. Two main methods in use include inductive and transductive TL.

With Inductive TL, tasks from the source domain differ from the tasks of the target domain. This method works with the presence or absence of labelled data. Multi-task learning is performed with labelled data while self-taught learning is performed without labelled data. A sub-category of inductive learning methods that focuses on applying the inductive approach to unsupervised learning tasks such as clustering, dimensionality reduction, and density estimation is referred to as unsupervised TL.

With Transductive TL, the tasks from both source and target domains are the same but the domains are different. This method is widely used in cross-lingual and domain adaptation TL studies.

Instance transfer, feature representation transfer, parameter transfer and relational knowledge transfer are approaches and all applied to inductive TL, while only instance transfer and feature representation transfer can be applied to transductive TL [2, 17, 27, 28].

Much of the success of TL is attributed to the use of Large Language Models (LLM) transformer models in general and failures to training time when using LLM that make the process computationally expensive [27]. The effectiveness of LLMs require that when training, the LLM is able to assimilate various learning points ranging from small spelling errors to the contextual meanings present in the training corpus. The TL method brings performance improvements when using LLM models because the models no longer have to be trained from scratch but can be used in a pre-trained state. Large crowd-sourced text corpora, such as the Stanford Question Answering Dataset (SQuAD) and Colossal Clean Crawled Corpus, are used to train LLMs such as BERT, RoBERTa, sBERT and Text-To-Text-Transfer-Transformers(T5) to enable them to perform several tasks [27]. The T5 model was trained using the Text-To-Text framework [15]. The T5 combines all NLP techniques such as translation, question answering, text summarisation, and document classification together in one model, thereby reducing the need to perform these tasks individually [20, 28].

3 Auto-generated Ontological Competency Questions (AgOCQs)

We describe the methods used to develop AgOCQs, and use as illustration its application to a small corpus of scientific articles on COVID-19.

As part of the development process, we leveraged the abstract representation from the linguistic pattern extraction method developed by [19] and the CLaRO, a controlled language set of templates for CQs by [3, 12]. These methods along with a combination that infuses NLP techniques and Transformer models make up the development processes for AgOCQs.

Fig. 1 and the algorithm in Fig. 2 summarise the design of the automated process of developing CQs, called AgOCQs. It begins by extracting domain text corpus which is then pre-processed using NLP techniques such as entity and sentence extraction, stop words removal [11] and regular expressions to produce cleaned data used as the input text data. We then apply the transductive TL where the source task for the model is the same as the target task. Using the pre-trained Text-to-Text-Transfer-Transformer (T5) base model [20], which is pre-trained with the SQUAD dataset to output a context, question and answer as our source task. We pass in our cleaned input data to undertake the same task as the source task of the base model; however, we only output the context texts and questions in this case. The corpus of questions generated from the input data is subsequently de-cluttered to remove duplicates and meaningless questions through a semantic grouping using the paraphrased algorithm of the Sentence Transformer model [21].

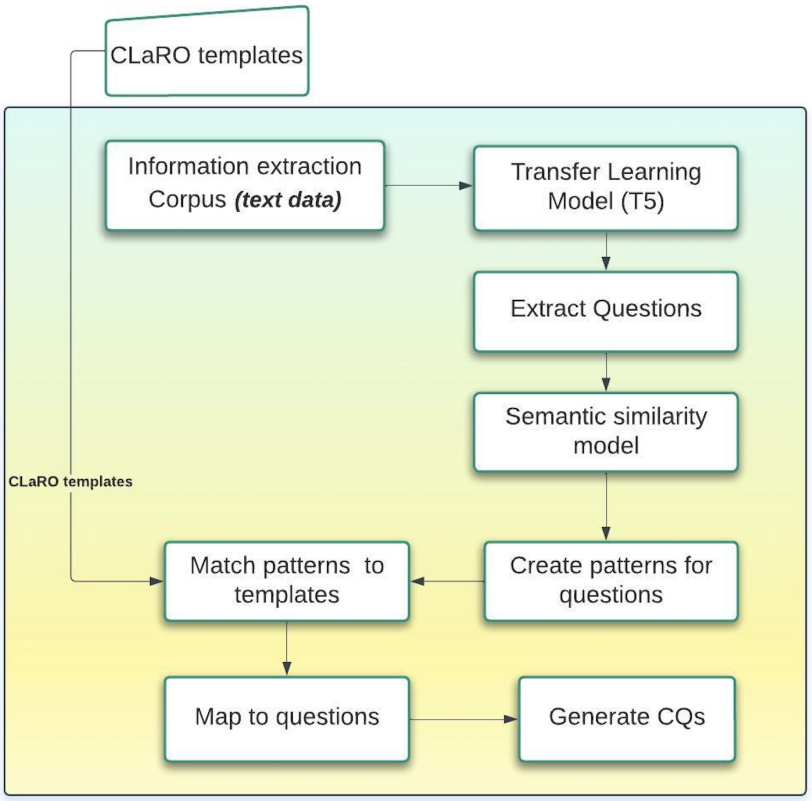

Fig. 1. Design of the pipeline architecture for automatic CQs authoring.

Next, we represent the questions into their linguistic abstract forms using the method from [19, 26]. To do this, each question is broken down into chunks and represented in the abstract form as *entity chunks(EC)* i.e., nouns/noun phrases, and *predicate chunks(PC)*, i.e., verbs/verb phrases, as illustrated in Table 1 for the use case chosen to test the approach with.

Generated questions in their abstract form are then compared to CQ templates from CLaRO that were also developed using the same abstract format [3, 12]. CLaRO templates, having been assessed for grammatical correctness and answerability by an ontology, serve as a gold standard for CQs in determining if a set of questions qualifies as CQs. If the abstract form of a question matches a template, the question is then referred to as a "Competency Question". The template rules are an integral part of the development process and help ensure that the abstract representation of the CQs produced correspond to the sentence chunk. The abstract forms can also be reviewed for any error as it is saved to a file that can be corrected and fed back into the process if needed.

```
input:  claroTemplates
output: modelBasedCQ, semanticClusteredTemplates
1. function informationExtraction(textCorpus):
    2. webScraping to download scientifc articles
   3. textExtraction to extract text sections from pdfs
   4. preProcessing to remove unwanted characters
      5. return cleanedText
6. function TransferLearningModel(SquadDataset, cleanedText)
      7. trainBase model for question generation
      8. modelEvaluation with cleanedText
         9. return ExtractQuestions
10. function similarityModel (ExtractQuestions)
            11. trainSimilarity model
            12. modelEvaluation model by cosine similarity
            13. return semanticClusters
14. function patternCreation(ExtractQuestions)
         15. chunkQuestions
         16. labelChunks as entity and predicate chunks
         17. return newCQsPatterns
18.  function templateMatching(newCQsPatterns, claroTemplates)
         19. comparePatterns to templates
         20. return matchedTemplates, unmatchedPatterns
21. function MapToQuestions(matchedTemplates, ExtractQuestions,
semanticClusters)
         22. findQuestions with corresponding templates
         23. groupQuestions with semantic clusters
         24. groupTempates to their semantic clusters
         25. return modelBasedCQs, semanticClusteredTemplates
26. main()
   27. informationExtraction (textCorpus)
   28. transferLearningModel(SquadDataset, cleanedText)
   29. similarityModel (ExtractQuestions)
   30. patternCreation(ExtractQuestions)
   31. templateMatching(newCQsPatterns, claroTemplates)
   32. MapToQuestions(matchedTemplates, ExtractQuestions, semanticClusters)
   33. return modelBasedCQ, semanticClusteredTemplates
```

Fig. 2. Outline of the main algorithm for automatic CQs authoring.

Table 1. Sampling of questions and their respective abstract representation, where the method as used on a small corpus of scientific articles on COVID-19.

Question	**Abstract form**
How many people have been infected with COVID-19?	How many EC1 PC1 been PC1 EC2?
What severity of the case may progress to respiratory distress or respiratory failure?	What EC1 of EC2 PC1 PC1 EC3 or EC4?
How severe is the disease related to age?	How PC1 is EC1 PC2 EC2?

Two groups emerge from this step in the pipeline: complete matches and variants (i.e., a very close match of a template). For the use case with the Covid-29 corpus (see below), only complete matches to templates were considered. With the CLaRO templates property of 1:m mapping, several questions can have abstract forms that correspond to just one of the templates. The abstract forms are then mapped back to the questions to give a set of CQs that are deemed ready for use by the ontology developer. A sample of the generated CQs for the use case and their corresponding sentence patterns mapping to templates are displayed in Table 3. All CQs can be found on Github: https://github.com/pymj/AgOCQs.

Table 2. CQs and corresponding sentence patterns, generated by the AgOCQs procedure when applied to the small Covid-19 text corpus

CQs	Templates	ID
How can Coronaviruses induce psychopathological sequelae?	How PC1 EC1 PC1 EC2?	17
What is the prevalence of emergent psychiatric conditions?	What is EC1 of EC2?	60
What is the mean age range for COVID19 survivors?	What is EC1 for EC2?	38
What is the role of SARSCOV2 immuneescape mechanisms?	What is EC1 of EC2?	60
What are the mainstay of clinical treatment?	What are EC1 of EC2?	60a
What is lymphopenia?	What is EC1?	90
What is the name for the cytokine storm14?	'What is EC1 for EC2?	38
What is a potential target for IL1 IL17?	What is EC1 for EC2?	38
What is another approach to alleviate COVID19 related immunopathology?	What is EC1 PC1 EC2?	66
What is the role of standardized treatment protocols for severe cases?	What is EC1 of EC2 for EC3?	61
What percentage of the subjects reported fatigue?	What EC1 of EC2 PC1 EC3?	68
What is the spread rate of COVID19?	What is EC1 of EC2?	60
What is the duration of symptoms for mild cases?	What is EC1 of EC2 for EC3?	61
What is a blood test for COVID19?	What is EC1 for EC2?	38
What role could corticosteroids play in severe cases?	What EC1 PC1 EC2 PC1 EC3?	58
What did the severe acute respiratory syndrome SARS attack reflect?	What PC1 EC1 PC1?	41

4 Evaluation of CQs generated with AgOCQs

The procedure described in Section 3 can be applied to any text corpus. The principal interest is obviously specialised subject domains, for which it is difficult to obtain extensive domain expert input. To this end, we created a small corpus of scientific articles on Covid-19, generated the questions with the proposed method, and subsequently evaluated them in a human evaluation. The evaluation approach, results, and discussion are described in the remainder of this section.

4.1 Approach

We conducted a survey to assess the CQs developed by AgOCQs. The aim of the survey was to use these target groups' responses on the answerability, relevance, scope, and grammaticality to assess the quality of these automatically generated CQs to determine how they fare on issues corresponding to some of the concerns associated with manually created CQs. The three groups are composed as follows: ontology experts, domain experts and ontology users. The participants had preliminary questions which were used to place them in the groups. The ontology users group is made up of 1) ontology professionals who do not consider themselves as experts. 2) ontology experts who are not the target domain experts. Thus, some participants in the users' group also appear in the ontology expert group.

The test corpus was created from freely available published research articles in the Covid-19 domain on the Web, in the time frame between 2020-2021. No specific criteria was applied other than the articles being a research paper from the COVID-19 domain. These articles were scraped using the PyPaperBot python tool. Seven articles only were used due to issues around compute capacity and lengthy processing time. The text corpus was created from the scientific articles' Introduction, Related work, and Methodology sections.

This corpus was then used in AgOCQs to generate candidate CQs for evaluation. Answerability, relevance, scope, and grammaticality are considered criteria for the assessment of the CQs. The target groups for the survey were Covid-19 domain experts, ontology experts, and ontology users. Users, in this case, were considered as research students working in areas directly or related to ontologies, and ontology experts working in the Covid-19 domain.

For CQs to be answerable, we presume that it should also be grammatically correct, therefore in assessing answerability we include the assessment of its grammaticality from participants with advanced English grammar competence. In terms of relevance, though the corpus used to develop the CQs is domain-specific, we still evaluate relevance as seen by domain and ontology experts as well as by ontology user groups. For the scope, all participants are asked to assess the overall CQs presented in the survey on their clarity of purpose from their standpoint. The objectives for the survey are, for each of domain experts, ontology experts, and ontology users as separate groups respectively:

1. To understand the respective judgments of the different target groups on the grammaticality and answerability of the automatically generated CQs.

2. To determine how the different groups rate the relevance of the automatically generated CQs.
3. To understand how each of the groups judge the automatic CQs as an indicator of the scope for their respective objectives.

For the analysis of the survey, we focus on the target groups' responses individually. We will analyze the overlaps of interests (if any) of the three target groups (ontology experts, Covid-19 subject experts, ontology users) and how that affects their judgments of CQs. We will also analyze how the experience of participants within each group could lead to different judgments. The survey contained question for classifying participants into different target groups as well as 20 CQs to be judged by the participants. A question was marked as a CQ as attention check, being CQ.17: *What is post-COVID similar to post-SARS syndrome?*) which serve to assist to assess some of the responses of the participants to CQs, since it is incoherent and should be judged accordingly. The last question was directed to how the participants judge the overall CQs in terms of coverage of the domain in question. A sampling of the CQs used the survey are shown in Table 3.

Table 3. A selection of the CQs used in the survey

CQ1: How can Coronaviruses induce psychopathological sequelae?
CQ2: What is the mean age range for COVID-19 survivors?
CQ4: What is the duration of symptoms for mild cases?
CQ5: What is the prevalence of emergent psychiatric disorders?
CQ7: What is lymphopenia?
CQ8: What is the current status of herd immunity?

To check for reliability, we apply Fleiss's Kappa's coefficient for inter-rater testing [7, 9] which measure to underscore the results from the target groups and remove any notion of occurrence based on chance. For interpretation, we use the scale from [14], where a moderate agreement to a near-perfect agreement would mean that the participants engaged with CQs and made an informed decision in their judgment. A widely used scale for measuring the agreement between raters is as follows: Zero (0) as No agreement, 0.1 - 0.20 as Slight, 0.21 - 0.40 as Fair, 0.41 - 0.60 as Moderate, 0.61 - 0.80 as Substantial, and 0.81 - 1.00 as Near perfect agreement.

Data and results are be available at https://github.com/pymj/AgOCQs.

4.2 Results

The results of the survey is presented on the basis of the assessment criteria of answerability, grammaticality, relevance, and scope in relation to the target

groups. There were 20 CQs in the survey to be judged, and a total of 17 respondents completed the survey. The domain expert group only had 1 respondent, the ontology expert group had 16 respondents, and the ontology users group (i.e., a combination of non-domain experts and non-ontology experts) had 15.

On grammatical competence and answerability, participants were asked to rate their English grammar competence (either average or very good). 75% rated their competence as *very good* while 25% rated themselves as *average*. Looking at the results by CQ from the participants that considered themselves to have very good English grammar competence, 70%-100% judged the majority of the CQs (14 of 20) as being grammatically correct Fig 3.

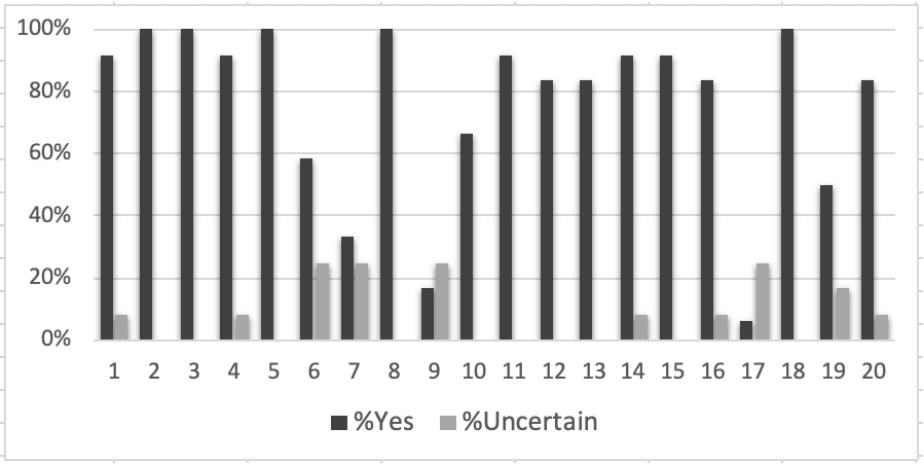

Fig. 3. Grammatical correctness by CQ

Looking at answerability on individual CQs from the same group on English grammar competence, the results showed that 50%-85% of participants deemed answerability to be positive in 12 CQs (see Fig 4).

Participants were also classified in terms of their competency in ontologies and CQs knowledge, where 81% identified as experienced and 19% as not experienced. Taking an overall view by CQs by experienced category, we observe answerability to be at an average of 50% and participants' uncertainty of answerability to be equally the same at 50% (see Fig 5).

In terms of relevance to the domain, 94% of participants classified themselves as not being experienced in the COVID-19 domain. As a result, the relevance of CQs was analyzed from the user's target group alone, which is composed of research students and ontology experts inexperienced in the COVID-19 domain. 70% of this group judged the CQs to be relevant to the domain Fig 6. Our results also showed that 69% of ontology and CQs experts who considered themselves experienced and 73% of ontology users believed the CQs gave them a clear scope of the Covid-19 domain (see Fig 7).

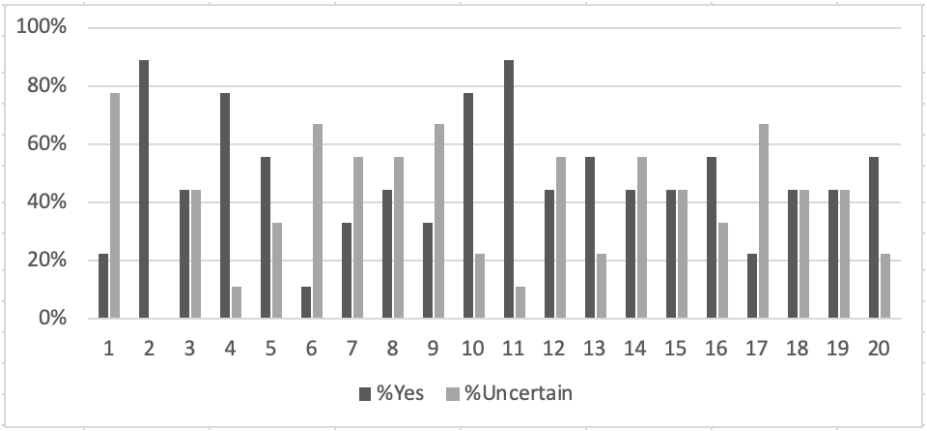

Fig. 4. Answerability by English grammar, ontology and CQs competence

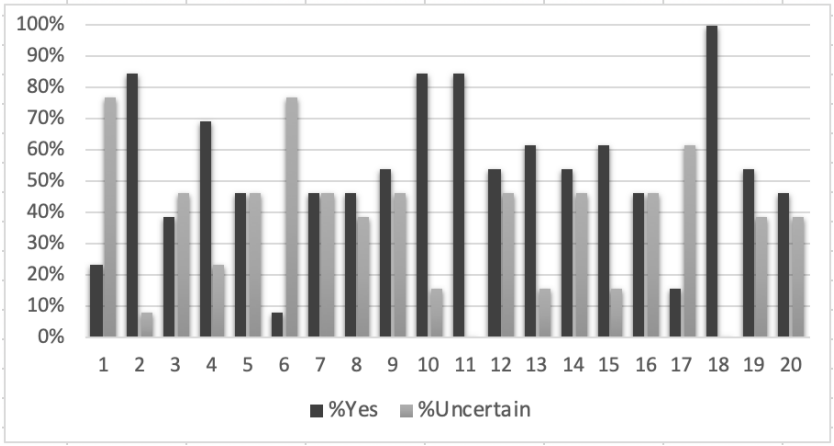

Fig. 5. Answerability by ontology and CQs experts

To ensure that the results from our survey represent the authentic view of our participants on key criteria used and did not occur by chance, we placed *CQ.17* as an attention check. Our results showed the participants passed the attention check overwhelmingly, with 95% of them detecting the CQ to be unanswerable with poor grammar. Also, we conducted a reliability test using Fleiss's Kappa coefficient test [14]. We interpret our results based on the scale from its Wikipedia page. Our Fleiss's Kappa scores for grammatical correctness, answerability and relevance were 0.55, 0.55 and 0.77, respectively. These scores show that the responses for the survey had a moderate to a substantial degree of agreement on the judgments made by participants, thus removing any notions that the responses occurred by chance.

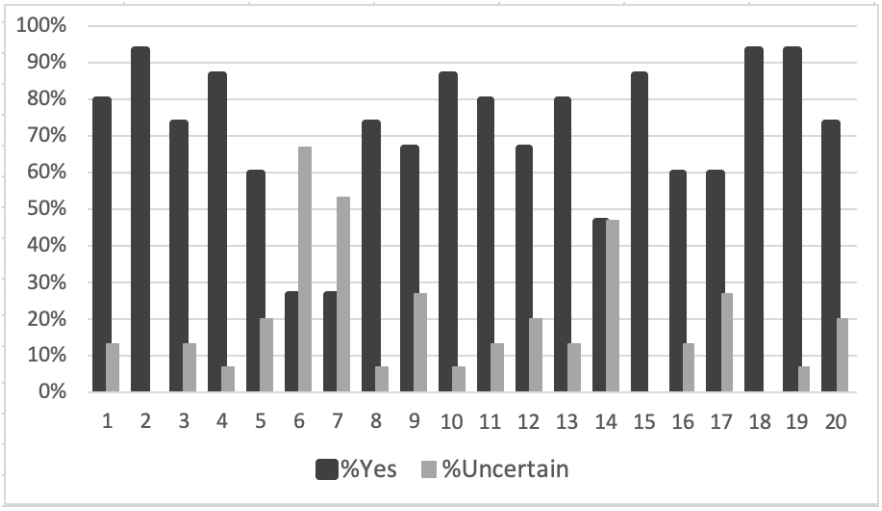

Fig. 6. Relevance to domain by ontology user group

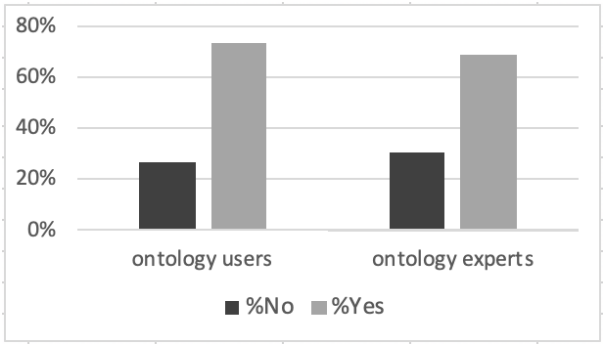

Fig. 7. Scope assessment by experts and user groups

4.3 Discussion

As mentioned in the motivation for this research, ontology developers have indicated that the process of manual CQ development is tedious and time-consuming and produces relatively few CQs of varying quality [26, 3, 12] which adds to the workload of ontology development and is a major reason why many skip the use of CQs. The proposed automated process AgOCQs resolves this problem by reducing the efforts of manual CQ authoring. This may potentially increase the adoption of CQs in ontology engineering processes. Compared to their manually curated counterparts, CQs from AgOCQs are granular, providing a large pool of CQs that are more specific compared to high-level scoping CQs. In addition, since most of our user target group also judged the CQs to cover relevant information on the domain Fig 7, suggests that AgOCQs may facilitate CQ reuse

across ontologies within the same domain or subdomain. The use of a text corpus in CQs development adds some of the benefits observed elsewhere [16, 24] to ontologies via CQs, thereby answering our first research question.

The results of the survey also provide a window to assess human judgment on CQs from ontology experts and engaged users. AgOCQs uses a set of templates considered as ground truth because the CQ templates have been verified to be answerable by an ontology where the contents are available. Yet, there is a divide in the results from experts on the certainty of the CQs being answerable compared to novices. Also, which is not uncommon for human judgements with a limited pool of participants with varied backgrounds, it shows that the opinions of ontology experts on grammaticality and answerability do not even out. For instance, although most participants indicated having very good English grammar competence (see Fig 3), there are CQs where the judgment by these participants is questionable; e.g., CQ.9: *What severity of the case may progress to respiratory distress?* is arguably grammatically correct when taken in context, but only 13% of the participants judged it as correct.

The automated CQs also face similar issues encountered by CQs from the manual process, especially on the criteria of answerability and grammatical correctness, as can be interpreted in agreement levels from the Fleiss Kappa scores, albeit to a lesser extent.

With CLaRO templates having a property of 1:n mapping to SPARQL queries and possible axiom patterns [3, 12], their use as part of the development of AgOCQs brings the possibility of realizing CQs reusability and by extension, ontology reusability within reach as This methodology can be applied to other domains with little or no reservations. However, new domain-specific templates which may currently not exist in CLaRO may cause the omission of certain questions.

A limitations of the study include the use of a small dataset to demonstrate the functionality of our method, which was due to the compute capacity available. Thus, the data used in this study is not representative of the data within the domain. We are working towards optimising the approach of the text processing as well as the use of an ontology to demonstrate its effectiveness on the issue of completeness in the ontology engineering process.

5 Conclusion and future work

The paper proposed AgOCQs to automate the process of creating competency questions for ontology development and selection. The results showed that CQs from AgOCQs using a domain text corpus are highly granular and provide a larger number compared to those manually developed. The evaluation indicated that it is possible to have a set of CQs that may serve a number of ontologies in the same domain. The proposed automated CQ creation process may foster CQ uptake for ontology selection, design, and evaluation. In future work, we plan to explore the effect of corpus size and genre on CQ generation.

Acknowledgements This work was financially supported by Hasso Plattner Institute for Digital Engineering through the HPI Research School at UCT. The authors also thank the survey participants for their participation.

References

1. Abdelghany, A.S., Darwish, N.R., Hefni, H.A.: An agile methodology for ontology development. International Journal of Intelligent Engineering and Systems **12**(2), 170–181 (2019)
2. Agarwal, N., Sondhi, A., Chopra, K., Singh, G.: Transfer learning: Survey and classification. Smart Innovations in Communication and Computational Sciences: Proceedings of ICSICCS 2020 pp. 145–155 (2021)
3. Antia, M.J., Keet, C.M.: Assessing and enhancing bottom-up CNL design for competency questions for ontologies. In: Proceedings of the Seventh International Workshop on Controlled Natural Language (CNL 2020/21). ACL, Amsterdam, Netherlands (2021), https://aclanthology.org/2021.cnl-1.11
4. Azzi, S., Iglewski, M., Nabelsi, V.: Competency questions for biomedical ontology reuse. Procedia Computer Science **160**, 362–368 (2019)
5. Bezerra, C., Freitas, F.: Verifying description logic ontologies based on competency questions and unit testing. In: ONTOBRAS. pp. 159–164 (2017)
6. Bezerra, C., Santana, F., Freitas, F.: Cqchecker: a tool to check ontologies in owl-dl using competency questions written in controlled natural language. Learning & Nonlinear Models **12**(2), 4 (2014)
7. Cyr, L., Francis, K.: Measures of clinical agreement for nominal and categorical data: the kappa coefficient. Computers in biology and medicine **22**(4), 239–246 (1992)
8. Dutta, B., DeBellis, M.: Codo: an ontology for collection and analysis of covid-19 data. arXiv preprint arXiv:2009.01210 (2020)
9. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological bulletin **76**(5), 378 (1971)
10. Haller, S.: Automatic Short Answer Grading using Text-to-Text Transfer Transformer Model. Master's thesis, University of Twente, the Netherlands (2020)
11. Honnibal, M., Montani, I.: spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing **7**(1), 411–420 (2017)
12. Keet, C.M., Mahlaza, Z., Antia, M.J.: CLaRO: a controlled language for authoring competency questions. In: Garoufallou, E., et al. (eds.) 13th Metadata and Semantics Research Conference (MTSR'19). CCIS, vol. 1057, pp. 3–15. Springer (2019), 28-31 Oct 2019, Rome, Italy
13. Keet, C.M., Lawrynowicz, A.: Test-driven development of ontologies. In: Sack, H., et al. (eds.) The Semantic Web. Latest Advances and New Domains - 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Proceedings. LNCS, vol. 9678, pp. 642–657. Springer (2016)
14. Landis, J.R., Koch, G.G.: An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics pp. 363–374 (1977)
15. Li, C., Su, Y., Liu, W.: Text-to-text generative adversarial networks. In: 2018 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2018)

16. Li, Q., Li, S., Zhang, S., Hu, J., Hu, J.: A review of text corpus-based tourism big data mining. Applied Sciences **9**(16), 3300 (2019)
17. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2009)
18. Peroni, S.: A simplified agile methodology for ontology development. In: Dragoni, M., Poveda-Villalón, M., Jiménez-Ruiz, E. (eds.) 13th OWL: - Experiences and Directions - Reasoner Evaluation - 13th InternationalWorkshop (OWLED'16). LNCS, vol. 10161, pp. 55–69. Springer (2016)
19. Potoniec, J., Wiśniewski, D., Ławrynowicz, A., Keet, C.M.: Dataset of ontology competency questions to SPARQL-OWL queries translations. Data in brief **29**, 105098 (2020)
20. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
21. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
22. Ren, Y., Parvizi, A., Mellish, C., Pan, J.Z., Van Deemter, K., Stevens, R.: Towards competency question-driven ontology authoring. In: The Semantic Web: Trends and Challenges: 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings 11. pp. 752–767. Springer (2014)
23. Suárez-Figueroa, M.C., Gómez-Pérez, A., Fernández-López, M.: The neon methodology for ontology engineering. In: Ontology engineering in a networked world, pp. 9–34. Springer (2012)
24. Tseng, Y.H., Ho, Z.P., Yang, K.S., Chen, C.C.: Mining term networks from text collections for crime investigation. Expert Systems with Applications **39**(11), 10082–10090 (2012)
25. Uschold, M., Gruninger, M.: Ontologies: Principles, methods and applications. The knowledge engineering review **11**(2), 93–136 (1996)
26. Wiśniewski, D., Potoniec, J., Ławrynowicz, A., Keet, C.M.: Analysis of ontology competency questions and their formalizations in sparql-owl. Journal of Web Semantics **59**, 100534 (2019)
27. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proceedings of the IEEE **109**(1), 43–76 (2020)
28. Zolotareva, E., Tashu, T.M., Horváth, T.: Abstractive text summarization using transfer learning. In: ITAT. pp. 75–80 (2020)