Evaluations of User-Driven Ontology Summarization

Conference or Workshop Item

Evaluations of User-Driven Ontology Summarization

Ning Li and Enrico Motta

Knowledge Media Institute
The Open University
Milton Keynes, United Kingdom
{n.li, e.motta}@open.ac.uk

Abstract. Ontology Summarization has been found useful to facilitate ontology engineering tasks in a number of different ways. Recently, it has been recognised as a means to facilitate ontology understanding and then support tasks like ontology reuse in ontology construction. Among the works in literature, not only distinctive methods are used to summarize ontology, also different measures are deployed to evaluate the summarization results. Without a set of common evaluation measures in place, it is not possible to compare the performance and therefore judge the effectiveness of those summarization methods. In this paper, we investigate the applicability of the evaluation measures from ontology evaluation and summary evaluation domain for ontology summary evaluation. Based on those measures, we evaluate the performances of the existing user-driven ontology summarization approaches.

Keywords: Ontology Summarization, Evaluation, Semantic Web

1. Introduction

Ontology Summarization, in recent years, has been recognised as an important tool, driven by users, to facilitate ontology understanding and help users quickly make sense of an ontology in order to support tasks like ontology reuse [1][2][3]. It has provided the basis for a number of user-centric technologies, such as the novel interactive frameworks for ontology visualization and navigation KC-Viz[1] and the online ontologies sharing and reusing system Cupboard[2]. Though there are a number of works in literature for ontology summarization, their evaluations are rather isolated from one another and there lacks a comparative view among the ontology summarization approaches. In fact, there lacks a systematic overview of what evaluation measures are there available and applicable for ontology summary evaluation.

As stated in [4], a key factor that makes a particular discipline or approach scientific is the ability to evaluate and compare the ideas within the area. For ontology summarization, evaluation measures can be very different depending on what drives or motivates the ontology summarization. For example, for task-driven ontology summarization, the evaluation can be task-specific and objective in that the criterion

[1] http://neon-toolkit.org/wiki/KC-Viz
[2] http://kmi-web06.open.ac.uk:8081/cupboard/

can be based on whether the ontology summary satisfies the requirements of the specific task as the original ontology does while gaining the expected benefits such as reduced ontology size. Whereas for user-driven ontology summarization, the ultimate goal is to serve the user with the help of ontology engineering tools, and therefore the evaluation can be subjective as well as objective. In this paper, we focus on the evaluation of user-driven ontology summarization, which has not been systematically addressed in literature. The main contributions of this paper consist in the following two aspects, one is to provide a systematic view of the evaluation measures for ontology summary with focus on user-driven ontology summarization, and the other is to evaluate two user-driven ontology summarization systems in a comparative way using evaluation measures investigated here.

This paper is organized as follows. In section 2, we provide an introduction to ontology summarization and a review of the ontology summarization works in literature in a categorical view. Following this, we focus on the introduction of user-driven ontology summarization works. Section 3 presents an investigation of evaluation measures and their applicability in ontology summary evaluation. In Section 4, we give our evaluation results in a comparative way of the two user-driven ontology summarization approaches. Section 5 concludes the paper with discussions.

2. Ontology Summarization

2.1 Ontology Summarization Overview

Motivated by the definition of text summarization in natural language processing, the authors in [2] provided a definition for ontology summarization as "the process of distilling knowledge from ontology to produce an abridged version for a particular user (or users) and task (or tasks)". According to this definition, the information content of a summary depends on either user's needs or/and task's requirements. This is one way of classifying the ontology summarization works in literature. In following, we provide a categorical view of the ontology summarization approaches.

Driving force. Ontology summarization is mostly driven by certain needs of users or tasks that are observable to ontology engineers. Task-driven ontology summarization works include winnowing ontology from very large size to the size only necessary to meet the needs of querying tasks [5], downsizing Abox to improve the scalability of ontology reasoning tasks [6]. User-driven ontology summarization focuses on the needs of users to understand and make sense of ontology quickly in large-scale ontology spaces, such as the work in [1][2][3] which will be the focus of this paper and will be given more details in Section 2.2.

Working unit of summarization. Driven by different motivations, ontology summarization can operate at different levels and on different constitutional components of ontology. For example, in [6], the summarization object is ABox with aims to improve the scalability of reasoning tasks for ontologies containing large ABoxes. The work in [2][3] operates on RDF sentence in order to have summaries which contain both terms (concepts and properties) and relations among the terms. Whereas in [1], summaries contain only concepts with the belief that they are more effective in just helping user to know what this ontology is about than further details like how the ontology is structured which may have an adverse impact of confusing inexperienced users, and also with the observation that further navigation of more details of the whole ontology from key concepts is feasible and practical.

Extractive or abstractive. For text in natural language, summarization can be extractive in that summaries are produced by selecting a subset of the elements in the original document, or abstractive by rephrasing the information content of the original document [7]. Although summaries produced by humans are typically not extractive, most of the scientific researches on summarization, notably text summarization, are on extractive summarization because abstractive summarization is much harder to implement due to problems of semantic representation, inference and natural language generation. Though some problems in text summarization like semantic representation, inference are no longer difficult to solve in ontology summarization, most of the ontology summarization works in literature are also extractive because, unlike text which is natural language that can be expressed in many different ways, ontology is a formalized representation of knowledge having a much simpler but stricter syntax and semantics. Also, Ontology is already a carefully selected, by domain experts, bunch of concepts as well as their relations, properties and facts.

2.2 User-driven Ontology Summarization

Contrary to task-driven ontology summarization, the ultimate goal for user-driven ontology summarization is to satisfy user's needs, which can be a very subjective matter. Therefore, it is not so easy to define a clear boundary of whether one summary is getting the job done or done better than another unless some clearly specified and commonly agreed criteria are laid. Therefore in following sections, we focus on the introduction of user-driven ontology summarization techniques, which will lead to the systematic investigation of ontology summary evaluation measures.

We have referred to three pieces of works, and the only three to the best of our knowledge, for user-driven ontology summarization aiming to facilitate ontology understanding. The work in [1] extracts key concepts as the best representatives of ontology and hence as ontology summary. In [1], a number of criteria were jointly considered, and correspondingly a number of algorithms were developed and linearly combined, to identify key concepts of an ontology. The criteria include: *name simplicity* which favors concepts that are labeled with simple names while penalizing compounds; *basic level* which measures how "central" a concept is in the taxonomy of the ontology; *density* highlights concepts which are richly characterized with properties and taxonomic relationships; *coverage* aims to ensure that no important part of the ontology is neglected; and *popularity* indentifies concepts that are commonly used. The summarization results, i.e. key concepts, were evaluated against human assessors' summaries, referred to as "ground truth". A good agreement has been found between algorithm-generated summaries with "ground truth".

There are other two works in [2] and [3] which also look into ontology summarization to facilitate user quickly make sense of what an ontology is about, but do not have the intention to support further exploitation of the ontology once user gets interested. Different from the work in [1], in [2] and [3], the authors take RDF sentence as the basic unit for summarization and extract the most salient/important sentences as summarization results. By constructing an RDF sentence graph with RDF sentences as vertices and links among them as edges, the authors calculate, for each vertex, a "centrality" value that determines the relative importance of a vertex within the graph. This work is largely motivated by the work of a graph-based text summarization [7]. Therefore, when it comes to evaluation, a lot of lessons have been learnt from the evaluation of text summarisation, which will be introduced with more details in Section 3.2.

3. Ontology Summary Evaluation

With those works around to do ontology summarization aiming to facilitate ontology understanding, it is more important than intriguing to compare their performances using similar, if not the same, evaluation measures in order to assist users in deciding the suitability of each approach to their own purpose. When approaching ontology summary evaluation, experiences have been gained from the two topics ontology summarization tries to cover or combine, apparently one is ontology evaluation, and the other is summary evaluation.

3.1 Ontology Evaluation

Ontology evaluation has been continuously researched since the beginning of ontology-supported engineering and the semantic web. However, there has not been a published work on ontology summary evaluation. There are similarities and dissimilarities between these two topics and therefore some, if not all, of the ontology evaluation approaches may be applicable for ontology summary evaluation. The ontology evaluation has been surveyed in work [4] and [8] and concisely summarized in [9]. Basically, the evaluation can be done automatically or manually and the evaluation can be carried out at different evaluation levels, which refer to the aspects of the ontology that are evaluated. A majority of the work in literature focus on the following three levels. 1) **Application-driven ontology evaluation**, in which the quality of an ontology is directly proportional to the performance of an application that uses it, 2) **Gold Standard based ontology evaluation**, where the quality of the ontology is expressed by its similarity to a manually built Gold Standard ontology, 3) **Corpus coverage ontology evaluation**, in which the quality of the ontology is represented by its appropriateness to cover the topic of a corpus.

Since ontology summarization aims to capture the most important parts of ontology while maintaining the focus of the conceptualized knowledge domain, theoretically, the above three types of evaluation schemes could be applied to ontology summary evaluation in a similar way as to ontology evaluation. However, as pointed out in [9], it has been recognized that they all have various levels of barriers to overcome when fulfilling ontology evaluation tasks. When it comes to ontology summary evaluation, these barriers actually become less prominent attributed fundamentally to the fact that the comparison now is not between two ontologies which most probably contain many different lexicons as well as concepts structured in many different ways, but between the original ontology and a subset of it. For example, in **Application-driven ontology evaluation**, problems lie in (a) the difficulty of assessing the quality of the supported task (e.g. search) and (b) creating a "clean " experimental environment where no other factors but the ontology influences the performance of the application [9]. When this type of evaluation is applied to ontology summary evaluation, these problems are no longer problems if the ontology summarization process is guided by the application, such as in [5] where the summarization is guided by the queries received from any dependent applications. When the guiding applications are then used to evaluate the resulted summary to check, for example, whether it can answer all the queries as the original ontology does, whether it can reduce the query-answering time etc. It tends to be straightforward and much easier than evaluating the original ontology. In **Gold Standard based ontology evaluation**, one of the difficulties encountered is that comparing two ontologies is rather difficult. According to [10], one of the few works on measuring the similarity between ontologies, ontologies can be compared at two different levels: lexical and conceptual. Lexical comparison assesses the similarity between the lexicons (set of labels denoting concepts) of the two ontologies. At the conceptual level the taxonomic structures and the relations in the ontologies are compared. These problems become easier in ontology summary evaluation, as seen in [1][2], because the comparison now is between candidate summaries with "gold standard" summary of the same ontology, also known as human assessors' "ground truth". Not only the size of ontology summary is a lot smaller than original ontology, also they are all confined to conceptual knowledge of the same domain. The same applies to the **Corpus coverage ontology evaluation**, though it has not been practiced in literatures for ontology summary evaluation. In the case of ontology summary evaluation, the topic coverage check is between candidate summaries and the original ontology and among candidate summaries without concerning how well the original ontology covers the topic in general and therefore much easier to implement. The applicability of ontology evaluation schemes for ontology summary evaluation and their practices in related work are listed in Table 1.

Table 1. Ontology evaluation schemes for ontology summary evaluation

Ontology evaluation Schemes	**Applicable for ontology summary evaluation**	**Practiced in**
Application-driven	Yes	[5]
Gold Standard based	Yes	[1][2]
Corpus coverage	Yes	N/A

3.2 Summary Evaluation

Just as experiences can be and has been gained from text summarization to do ontology summarization, lessons can be learned from the evaluation of text summarization to do the evaluation of ontology summarization. This happens in the work of [2] where graph-based text summarization techniques were applied in ontology summarization. The authors evaluated the summary quality as well as the effectiveness of ontology summarization techniques using measures similar to the evaluation of text summarization. According to [11], many evaluation measures proposed for text summarization can be classified into three categories: **Recall-based**, **Sentence-rank-based** and **Content-based**. In general, one measure will produce one score for each summary produced by one technique. If summaries produced by one technique consistently produces higher scores by all measures than those produced by other techniques and thus has a higher average score, it is reasonable to believe that the summarization technique that produces the summaries with higher scores is a better technique.

Recall-based measures rely on human assessors to extract "ground truth" summaries first and then compare machine-generated summaries with them by counting the number of sentences they have in common. Therefore, the more sentences a summary has *recalled* from the "ground truth", the higher the evaluation score will be. Though intuitive and simple, recall-based evaluation measures fall short in a number of aspects. Firstly, they introduce a bias because human assessors are used and, in extracting sentences of a document, the agreement among assessors is typically quite low [12]. Secondly, a small change in the summary output (by replacing one sentence with another equally good one which happens not to match majority "ground truths") could change the evaluation score dramatically. This measure is practised in [1] between machine-generated summaries with "ground truth". **Sentence-rank-based** evaluation measures also rely on human assessors to produce "ground truth". However, instead of producing summaries that simply contain the most representative sentences of a document, each assessor is asked to rank the sentences of a document in the order of their importance in the summary. All machine-generated summaries also contain the rankings of the sentences. Kendall's *tau* statistic [13] can then be used to quantify a summary's agreement with a particular "ground truth". This *tau* measure is used in [2] to compare the performances of different "centrality"-based summarization methods among themselves and with "ground truth". **Content-based** measures, as the name indicates, calculate the similarity of the summary content to the full document content. The content similarity can be simply realized by finding the "term frequency" vectors of summaries and those of documents, and then calculating inner product value of two vectors which will result in a score indicating how similar the two vectors are [11]. In practise, "ground truth" can also be used as an alternative to the full document, though not a necessity as in the two measures mentioned above. That is to say, if "ground truth" is available, the similarity comparison can be carried out against the "ground truth" instead of the full document. This is what was practised in [2], where "vocabulary overlap" is used instead of "term frequency" to measure the similarity between machine-generated summaries with "ground truth". In this context, recall-based measures can be loosely regarded as a special case of content-based measures. The applicability of ontology evaluation schemes for ontology summarization evaluation and their practices in related work are listed in Table 2.

Table 2. Summary evaluation scheme for ontology summary evaluation

Summary evaluation schemes	Applicable for ontology summary evaluation	Practiced in
Recall-based	Yes	[1]
Sentence rank based	Yes	[2]
Content-based	Yes	[1][2]

4. User-driven Ontology Summary Evaluation: A Comparative Discussion

So far, the few works as explained in Section 2.2, which are user-driven and have focus on ontology summarization to facilitate ontology understanding remain unrelated and have not been evaluated on a common ground. Therefore, it is hard for readers to infer how well each system performs in a comparative way. As described in Section 3 that, among the only few, the works in [1] and [2] deploy measures compliant with those from both ontology evaluation and summary evaluation to evaluate their summarization results. In specific, they both deploy **Gold Standard based** and **Content-based** evaluation measures. To be specific, they both rely on human assessors' "ground truth" contents, with which the summarization results are compared by content. Given this commonality, in this section, we will present the evaluation of the two works in a comparative way by setting up a common ground for comparison. Not only will it serve as a test case for our claims that ontology summarization, like any other scientific work, can and should be evaluated by similar, or the same, measures, also it provides an indicative view of how state-of-the-art user-driven ontology summarization approaches perform and hence allows users to decide the suitability of each approach to their own purpose. The evaluation in [3] is excluded here because it could not rely on human assessors' "ground truth" information and hence lack the common ground with the other two approaches.

In the work of [2], the basic unit of summarization is RDF sentence, which is, in a nutshell, an integrated information unit that can be a single RDF statement without any blank node or a set of RDF statements connected by blank nodes. Using RDF sentence as basic distilling unit instead of terms (namely concepts), the authors claimed that it would provide extra knowledge of how the terms are related in the ontology and therefore provide a more comprehensive understanding of the ontology. However, in their work, not all, but only "generic" RDF sentences are considered in ontology summary while excluding "special" sentences. The "generic" there referred to RDF sentences which are mapped from axioms of atomic class and property whereas the "special" referred to sentences containing axioms specifying equivalence between class restrictions for example. However, in the work of [1], the basic working unit for summarization is concepts only. This is due to the fact that this work was mainly driven by a foreseen use case scenario, that is to facilitate the graph-based visualisation and navigation for large-scale ontologies. For ontology visualisation tools, concepts are the most fundamental component and play the most important role which interlinked with each other with the links among them being either properties or axioms of classes, such as *isA* relations. When ignoring the links between atomic classes and ignoring the atomic properties in the results of [2], we obtain a set of result which contains concepts only as the results in [1]. The results in these two works can therefore be evaluated comparatively by jointly applying **Gold Standard based** and **Content-based** measures.

In our evaluation, we use two test onotlogies *biosphere*[3], *financial*[4] which have been used in [1] as test ontologies because they have much larger vocabularies than those used in [2], and contain no properties that are subject to extraction by the summarization approaches in [2] hence enhancing the comparability of these two works. Using approaches in [2] with their online services[5], we obtain the most salient RDF sentences, from which we extract the first 20 concepts as key concepts in their order of appearance in the summarised RDF sentences by ignoring the links among them. We then calculate the number of key concepts that agree with "ground truth", i.e. "gold standard", that is the majority, over 50%, opinion of the human assessors to reduce the versatile subjectivity of individual assessors. Eight assessors with good experience in ontology engineering were asked to extract 20 concepts they considered the most representatives for each ontology. The statistics of the test ontologies, the number of key concepts agreed by more than 50% of human assessors and the number of key concepts extracted by those two summarization approaches are listed in Table 3. Table 4 and Table 5 list the 20 key concepts extracted by the two summarization approaches respectively with key concepts that agree with respective "ground truth" highlighted using italic font. The "ground truth" of *biosphere* ontology include *Animal, Bird, Fungi, Insect, Mammal, MarineAnimal, Microbiota, Plant, Reptile, Vegetation*, and that of *financial* ontology include *Bank, Bond, Broker, Capital, Contract, Dealer, Financial_Market, Order, Stock*.

Table 3. Statistics of ontology and ontology summarization results

Ontology	**No. of concepts**	**No. of concepts in "ground truth"**	**No. of key concepts extracted by approach in [1]**	**No. of key concepts extracted by approach in [2]**
biosphere	87	10	8	6
financial	188	9	6	6

Table 4. Key concepts extracted by summarization approach in [1]

Ontology	**Key Concepts**
biosphere	*Animal, Bacteria, Bird*, Crown, Fish, *Fungi*, FungyTaxonomy, Human, Litter, LivingThing, *Mammal, MarineAnimal*, Marine- Plant, *Microbiota*, MicrobiotaTaxonomy, Mold, Mushroom, *Plant, Vegetation*, Yeast
financial	Agent, *Bond, Capital*, Card, Cost, *Dealer*, Financial_Asset, Fi-nancial_Instrument, *Financial_Market*, Money, *Order*, Organization, Payment, Price, Quality, Security, *Stock*, Supplier, Transaction, Value

[3] http://sweet.jpl.nasa.gov/ontology/bioshpere.owl
[4] http://www.larflast.bas.bg/ontology
[5] http://iws.seu.edu.cn/services/falcon-f/ontosum/

Table 5. Key concepts extracted by summarization approaches in [2]

Ontology	Key Concepts
biosphere	*Plant*, LivingThing, *Vegetation*, *Microbiota*, *Animal*, MicrobiotaTaxonomy, *Fungi*, FungiTaxonomy, MarinePlant, *Mammal*, Fish, Canopy, Macroalgae, Macroalgae, Anemone, Mushroom, Protist, BlueGreenAlgae, Foraminifer, VegetationCover
financial	Financial_instrument, financial_asset, security, *bond*, asset, agent, financial_agent, *organization*, *capital*, finicial_ mean, debt_instrument, *market*, financial_market, municipal_bond, market_agent, *stock*, *contract*, supplier, corporate_bond, government_bond

We can see from the results that, with respect to the total number of concepts in the test ontologies which are 87 and 188 respectively, there is a quite high correlation between the results from these two approaches. This means that though they work on different units of ontology and different methodologies are used to do ontology summarization, the results in terms of concepts are close. This is due to the fact that calculating the "centrality" of RDF sentences in [2] having similar effects with calculating *basic level*, *density* and *coverage* in [1] because they all work on the topology and taxonomy aspects of ontology. By comparing them in details, we can tell the influence of each approach on the final results. The results from [1] approach contain more simple names, i.e. non-compound names, and more popular names than those from [2] approach. This is not surprising because the latter approach works only on ontology structures whereas the earlier one works not only on ontology structures, but also on *name simplicity* and *lexical popularity*.

5. Discussions and Conclusions

This paper provides an overview of state-of-the-art ontology summarization approaches in a categorical view first, and then focuses on the introduction of user-driven ontology summarization approaches. With aims to evaluate user-driven ontology summarization approaches on a common ground, in this paper, we systematically investigate evaluation measures from both ontology evaluation and summary evaluation domains and inspect their applicability for ontology summary evaluation. After spotting the commonality between the evaluation measures used by the existing ontology summarization approaches. We provide our evaluation results on two of the user-driven ontology summarization approaches by looking at them together in a comparative way. This evaluation confirms the comparability of those ever-isolated approaches given a common ground and a same set of evaluation measures, and also provides an indicative view of how well each approach performs in comparison with each other.

User-driven ontology summarization has provided the basis for a number of user-centric technologies, such as KC-Viz and Cupboard, and also for experiments in *Cautious Knowledge Sharing* [14], where ontology providers only advertise ontologies through automatically generated summaries, rather than in their entirety. It has also been used to index ontologies repositories using key concepts only, as opposed to indexing ontologies using the totality of their concepts [14]. The results showed a slight degradation in performance but apparently with a significant decrease in the index size. This will simplify the deployment of semantic repositories in scenarios such as those envisaged by the SmartProducts project[6], where we aim to deploy semantic technologies in domestic devices with very limited RAM and computational power. Therefore, evaluations of ontology summary are of paramount importance to the development of effective summarization approaches to ensure their performances in different scenarios.

6. References

1. Peroni, S., Motta, E., d'Aquin, M.: Identifying Key Concepts in an Ontology Through the Integration of Cognitive Principles with Statistical and Topological Measures. In: 3rd Asian Semantic Web Conference, Bangkok, Thailand (2008)
2. Zhang, X., Cheng, G., Qu, Y.: Ontology Summarization Based on RDF Sentence Graph. In: 16th Inter. World Wide Web Conference Banff, Alberta, Canada, May 8-12 (2007)
3. Zhang, X., Cheng, G., Ge, W., Qu, Y.: Summarizing Vocabularies in the Global Semantic Web. Journal of Computer Science and Technology 24(1), pp.165-174. Jan. (2009)
4. Brank, J., Grobelnik, M., Mladenic., D.: A Survey of Ontology Evaluation Techniques. In: Conference on Data Mining and Data Warehouses (SiKDD), Ljubljana, Slovenia (2005)
5. Alani, H., Harris, S., O'Neil, B.: Winnowing Ontologies Based on Application Use. In: 3rd European Semantic Web Conference (ESWC), Budva, Montenegro, June (2006)
6. Fokoue, A., Kershenbaum, A., Ma, L., Schonberg, E., Srinivas, K.: The Summary Abox: Cutting Ontologies Down to Size. In: 5th International Semantic Web Conference (2006)
7. Erkan, G., Radev, D.R.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research, 22, pp. 457–479 (2004)
8. Hartmann, J., Sure, Y., Giboin, A., Maynard, D., Suarez-Figueroa, M.C., Cuel. R.: Methods for Ontology Evaluation. Knowledge Web Deliverable D1.2.3 (2005)
9. Sabou, M., Lopez, V., Motta, E., Uren, V.: Ontology Selection: Ontology Evaluation on the Real Semantic Web. In: Workshop: Evaluation of Ontologies for the Web (EON) at 15th International World Wide Web Conference, Edinburgh (2006)
10. Maedche, A., Staab, S.: Measuring Similarity between Ontologies. In: Int. Conf. on Information and Knowledge Management (CIKM). LNAI vol. 2473 (2002)
11. Donaway, R.L., Drummey, K.W., Mather, L.A.: A Comparison of Rankings Produced by Summarization Evaluation Measures. In: ANLP/NAACL Workshop on Automatic Summarization, pp 69–78 (2000)
12. Jing, H., McKeown, K., Barzilay, R., Elhadad, M.: Summarization Evaluation Methods: Experiments and Analysis. In American Association for Artificial Intelligence Spring Symposium Series, pp. 60-68 (1998)
13. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures. CRC Press (1997)
14. Lopez, V., Guidi, D., Motta, E., Peroni, S., d'Aquin, M., Gridinoc, L.: Evaluation of Semantic Web Applications. OpenKnowledge Project Deliverable, D8.5 (2008)

[6] http://www.smartproducts-project.eu/