A Collaborative Framework for Ontology and Instance Data Co-Evolution and Extraction

Omar Qawasmeh

Univ. Lyon, CNRS, Lab. Hubert Curien UMR 5516, F-42023 Saint-Étienne, France
omar.alqawasmeh@univ-st-etienne.fr

Abstract. Ontologies are at the heart of the semantic web, i.e. making data published on the web comprehensible to intelligent added value services. Ontologies consensual design ensures its usefulness and wide acceptance by service developers. The collaboration of ontology engineers, domain experts from multiple disciplines, and end-users is required to design, evolve, and populate ontologies with the instance data. In this thesis, we investigate in the different systems that are concerns in designing and evolving the ontologies in automatic or semi-automatic techniques. Moreover, we are working to provide a collaborative framework for ontology and instance data co-evolution and extraction.

Keywords: Ontology · Knowledge Engineering · Collaborative ontology development · Collaborative ontology evolution.

1 Introduction

As defined by Gruber [8], an ontology is an explicit formal specification of terms in some domain, and relations between those terms. Ontologies play nowadays an important role in organizing and categorizing data in information systems and on the web, which leads to a better understanding, sharing and analyzing of knowledge in a specific domain. Several methodologies have been proposed for the development process of ontologies (e.g. [9, 2, 12]). One of the earliest methodologies was proposed by Noy et al. [15], and consists of a set of phases to be followed by the knowledge engineers: 1. decide the domain and the usage of the ontology, 2. decide on existing ontologies to reuse, 3. define the class hierarchy along with their relations and properties, and 4. create instances based on the class hierarchy to populate the ontology.

As mentioned in [5], the development process of an ontology in a fully manual way can be a very complex task to achieve. This motivates the design and development of semi-automatic or fully-automatic tools to assist the knowledge engineer in the ontology development process. In this research we investigate in collaborative ontology and instance data co-evolution and extraction.

The rest of this paper is organized as follows. Section 2 presents the the problem statement and the research questions we introduce in this thesis. Then in Section 3 we propose the current state of the art in the field, we here focus only on the systems that are addressing the first research question. In Section 4 we present our approach for the design of a bootstrapping functionality in the ontology development process to answer the first research question. Section 5 reports on the results of experiments for evaluation our approach, and Section 6 conclude the paper and shows the future work.

2 Problem statement

We define the Collaborative ontology evolution as the evolution of the axioms or the entities of an ontology by taking into account the advantage of different resources. Resources can be either computer systems, other external ontologies, external knowledge bases, or domain experts. The general research idea that is proposed by this thesis is to provide a collaborative framework for ontology and instances co-evolution and extraction. Our working pipeline is divided into 3 research questions:

1. Studying the quality of different knowledge bases to improve the ontology development. In [16], we investigate the problem of ontology construction in both automatic and semi-automatic approaches. There are two key issues for the ontology construction process: the cold start problem (i.e. starting the development of an ontology from a blank page) and the lack of availability of domain experts. We describe a functionality for ontology construction based on the bootstrapping feature. For this feature, we take advantage of three large public knowledge bases: Dbpeida [3], Wikidata [18], and NELL [7]. We report on a comparative study between our system and the existing ones on the wine ontology[1].
2. Studying the impact of ontology evolution on other artifacts that relies on it. Linked Open Vocabulary (LOV) [17] is considered as a rich repository of ontologies (i.e. vocabularies). LOV's main goal is to help publishers and users of linked data and vocabularies to assess, reuse, and publish different vocabularies based on their needs. LOV currently has 648 different vocabularies[2], each one is described with different properties, such as number of incoming links (i.e. how many ontologies are using ontology x), number of outgoing links(i.e. how many ontologies are using by ontology x), number of different versions, and data-sets. The outcome of this step is a study that shows the main changes between the different ontology versions that actually affect the evolution process. This study is used later in order to enhance the development process of ontologies, which could help in enhancing the evolution of the ontologies later.
3. Solving the impact of ontology evolution using collaborative techniques which can help to enhance the quality of LOV. In this phase, we will provide a framework in order to facilitate the process of ontology evolution. We will rely on the outcomes from the previous steps to design the functionality of the new frame work.

[1] https://www.w3.org/TR/owl-guide/wine.rdf

[2] last check Jun 2018

3 Automatic Ontology Development: A State of the art

In this section we focus on studying the different approaches that are interesting to answer the first research question (i.e. Automatic Ontology Development). Bedini et al. [4] classifies the approaches for automatic ontology development into the following four categories:

1. **Conversion or translation:** approaches that use conversion or translation algorithms to construct ontologies from a well-defined representation such as XML or UML. This approach shows a high automation ratio, however, it does not really address the problem of ontology construction.
2. **Mining based:** approaches that use data mining or natural language processing algorithms to construct ontologies. These approaches process unstructured data or text. The approaches in this category require human assistance to help mine or organize the different concepts extracted from the data sources.
3. **External knowledge based:** approaches that use external knowledge bases to construct or to enrich the ontologies. Examples of such external knowledge bases include WordNet [13], Wikidata [18], DBpedia [3], etc.
4. **Frameworks:** approaches that integrate different modules to achieve the goal of constructing ontologies.

Our proposed approach follows the third category (External knowledge based), so in the next subsection, we shortly present some of the relevant approaches that are classified in the same category.

3.1 Ontology Development based on External Knowledge

Kong et al. [11] use WordNet [13] as a general ontology to extract a set of concepts to build a domain specific ontology. Their system queries WordNet based on a set of keywords to extend the ontology by adding the list of new concepts. They compare their results to the wine ontology[3] developed by W3C. Table 4 shows their results comparing to the wine ontology. Examples of other approaches that use WordNet as an external knowledge base include [14, 1].

Kietz et al. [10] propose an approach that uses three knowledge bases to construct ontologies. Each one of the knowledge bases is used to achieve a specific task. The three knowledge bases are: 1. a generic ontology to generate the main structure, 2. a dictionary containing generic terms close to the required domain, and 3. a textual corpus specific to the required domain to enhance and clean the ontology from unrelated concepts. The result is an ontology composed of 381 terms (200 new terms) and 184 relations (42 new relations). The new terms and relation is added to a baseline ontology.

Cahyani and Wasito [6] propose an automatic system to build an ontology for the Alzheimer's disease. Their system consists of the following steps: 1. term

[3] https://www.w3.org/TR/owl-guide/wine.rdf

relation extraction, 2. matching the relations to Alzheimer glossary [4], 3. matching with ontology design patterns, 4. similarity computation, and 5. ontology building and evaluation. To evaluate their system they use a list of 125 papers on Alzheimer disease. Their system is able to retrieve 1,995 correct terms with 42 relations.

After studying the approaches we mentioned earlier, we found that most of them make use of predefined dictionaries (e.g. list of concepts) or lexicons (e.g. WordNet), or they use specialized glossaries (e.g. Alzheimer glossary). Several limits can be listed regarding these resources: the existence and availability of such dictionary or glossary for a given domain, the limited richness of the vocabulary, and the supported languages (generally limited to English). Based on this analysis, we propose in the next section an original functionality for semi-automatic ontology development tools.

4 A semi-automatic Approach for Bootstrapping Ontology Development with External Knowledge Bases

In order to improve current automatic ontology construction, we propose a functionality using publicly available knowledge bases: DBpedia, Wikidata and NELL. The pros of using these knowledge bases are that they are structured (RDF for DBpedia and Wikidata; specific data format for NELL), very large, include rich relations, are dynamic (i.e. evolving in time), machine understandable and multilingual. We use DBpedia and Wikidata to gain information and generate a list of classes and relations, and we use NELL knowledge base to generate instances of these classes.

We follow a semi-automatic bootstrapping technique, where the user enters a set of keywords related to a specific domain (e.g. wine, grapes, wine color, wine region, for the wine domain). Then by issuing a series of queries to the external knowledge bases, several classes and relations are extracted. Then this generated list is shown to the user for selection. After the user's validation, the set of classes is used to extract the instances from the NELL knowledge base. Interested readers may refer to [16] for more details. The algorithm we used can be found in Algorithm 1.

5 Preliminary Results

Most approaches cannot be evaluated on an arbitrary domain as it would require numerous specific data sources: a specific database on a domain, a corpus describing the domain, existing ontologies on the domain, etc. So, in order to validate our approach, we compare our results to those published in [11]. Recall authors in [11] proposed an ontology construction approach based on WordNet, and validated it comparing the numbers of extracted classes, properties and instances with the W3C's wine ontology[5]. We therefore lead a similar experiment

[4] https://www.alz.org/care/alzheimers-dementia-glossary.asp

[5] https://www.w3.org/TR/owl-guide/wine.rdf

Algorithm 1: The General Algorithm Implemented by our System

```
ConstructInitialOntology($keywords$);
Input  : $keywords$, a list of keywords given by the domain expert
Output : $\langle classes, relations, instances \rangle$ lists of terms to bootstrap the
         ontology.
$\langle classes, relations, instances \rangle \leftarrow \langle \varnothing, \varnothing, \varnothing \rangle$
foreach $keyword$ in $keywords$ do
    $\langle abstract, labels, uri \rangle \leftarrow$ queryDBPedia($keyword$)
    $\langle classes, relations \rangle \leftarrow$ queryWikiData($keyword$)
    $instances \leftarrow$ queryNELL($keyword$)
    $\langle classes', relations', instances' \rangle \leftarrow$
      pick($abstract, labels, uri, classes, relations, instances$);   // let the user
      pick the terms he wants
    $classes \leftarrow classes \cup classes'$;
    $relations \leftarrow relations \cup relations'$;
    $instances \leftarrow instances \cup instances'$;
return $\langle classes, relations, instances \rangle$ ;
```

to evaluate our system, and we compare our results to the baseline ontology (the W3C's wine ontology) and to the results in [11]. Authors in [11] use keyword "wine" to perform a query over WordNet. So that the comparison it fair, we used the same keyword "wine" as an input to our system. The raw results of our experiment, i.e., the full lists of classes, relations, and instances, our system suggests to the user, are made available in a Google sheet online[6]. Table 4 gives an overview of these results are compare them to the W3C's wine ontology and to the results of [11]. Out of the 80 classes our system extracted, 11 were already part of the W3C's wine ontology. We judge the remaining 69 relevant for a Wine ontology, so they could be used to extend this existing ontology. Our system also extracted 6 relations as listed in Table 2, apart from instanceOf and subClassOf, all of them are relevant for a wine ontology but not in the set of relations the W3C's wine ontology declares. As for the instances, we extracted 500 instances from NELL using a confidence threshold of 0.94 [7] to filter NELL's beliefs. This experiment shows that our system performs better than [11] while proposing only relevant concepts, which allows us to assert it would be a good fit for the bootstrapping phase of ontology development.

6 Conclusion and Future work

In this paper we presented the research questions that is addressed by the thesis. Regarding the first research question, we made a thorough state of the art on automatic ontology construction approaches, and we proposed an approach for

[6] "wine" experiment: full lists of terms our System outputs http://bit.ly/2EEKItn

[7] The threshold value was chosen based on a set of experiments on a different set of keywords.

Approach	W3C's wine ontology	[11]'s wine ontology	Our Approach
Class Number	74	62	**80**
Property Number	**13**	7	6
Instance Number	161	98	**500**

Table 1: Comparison of the Number of Classes, Relations, and Instances between our proposed approach, [11]'s approach and the W3C's wine ontology

ontology bootstrapping based on the use of three external knowledge bases: DBpedia, WikiData, an NELL. Preliminary results shows that our system performs better than [11] that is based on WordNet, and proposes only relevant concepts. This allows us to assert it would be a good fit for the bootstrapping phase of ontology development, and could even be reused as a first step before applying other techniques.

Currently we are working on both of the second and third questions. An expected outcome is to have a global study of the evolution of the different ontologies that are included in Linked Open Vocabulary (LOV) knowledge base. This study will be used later in order to provide a set of recommendations to the knowledge engineers to enhance the development process of their ontologies.

Finally, we are going to provide a frame work that facilitate the development process of ontologies taking into account the collaborative features.

7 Acknowledgment

This thesis is supervised by Prof. Pierre Maret (Hubert Curien Laboratory, Université de Lyon), Dr. Maxime Lefrançois (École des Mines de Saint-Étienne), and Dr. Antoine Zimmermann (École des Mines de Saint-Étienne).

References

1. Agirre, E., Ansa, O., Hovy, E.H., Martínez, D.: Enriching very large ontologies using the WWW. In: ECAI'2000 Workshop on Ontology Learning, Proceedings of the First Workshop on Ontology Learning OL'2000, Berlin, Germany, August 25, 2000. Held in conjunction with the 14th European Conference on Artificial Intelligence ECAI'2000, Berlin, Germany (2000), http://ceur-ws.org/Vol-31/EAgirre_14.pdf
2. Apisakmontri, P., Nantajeewarawat, E., Buranarach, M., Ikeda, M.: Ontology Construction and Schema Merging Using an Application-Independent Ontology for Humanitarian Aid in Disaster Management, pp. 281–296. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-15615-6_21, https://doi.org/10.1007/978-3-319-15615-6_21
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: Dbpedia: A nucleus for a web of open data. In: The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007. pp. 722–735 (2007). https://doi.org/10.1007/978-3-540-76298-0_52, https://doi.org/10.1007/978-3-540-76298-0_52

4. Bedini, I., Nguyen, B.: Automatic ontology generation: State of the art. PRiSM Laboratory Technical Report. University of Versailles (2007)
5. Blomqvist, E.: Pattern ranking for semi-automatic ontology construction. In: Wainwright, R.L., Haddad, H. (eds.) Proceedings of the 2008 ACM Symposium on Applied Computing (SAC), Fortaleza, Ceara, Brazil, March 16-20, 2008. pp. 2248–2255. ACM (2008). https://doi.org/10.1145/1363686.1364224, http://doi.acm.org/10.1145/1363686.1364224
6. Cahyani, D.E., Wasito, I.: Automatic ontology construction using text corpora and ontology design patterns (odps) in alzheimer's disease. Jurnal Ilmu Komputer dan Informasi **10**(2), 59–66 (2017)
7. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E.R.H., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: Fox, M., Poole, D. (eds.) Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010. AAAI Press (2010), http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1879
8. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge acquisition **5**(2), 199–220 (1993)
9. Iqbal, R., Murad, M.A.A., Mustapha, A., Sharef, N.M.: An analysis of ontology engineering methodologies: A literature review. Research journal of applied sciences, engineering and technology **6**(16), 2993–3000 (2013)
10. Kietz, J.U., Maedche, A., Volz, R.: A method for semi-automatic ontology acquisition from a corporate intranet. In: EKAW-2000 Workshop "Ontologies and Text", Juan-Les-Pins, France, October 2000 (2000)
11. Kong, H., Hwang, M., Kim, P.: Design of the automatic ontology building system about the specific domain knowledge. In: Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference. vol. 2, pp. 4–pp. IEEE (2006)
12. Kotis, K., Vouros, G.A.: Human-centered ontology engineering: The HCOME methodology. Knowl. Inf. Syst. **10**(1), 109–131 (2006). https://doi.org/10.1007/s10115-005-0227-4, https://doi.org/10.1007/s10115-005-0227-4
13. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM **38**(11), 39–41 (1995). https://doi.org/10.1145/219717.219748, http://doi.acm.org/10.1145/219717.219748
14. Moldovan, D.I., Girju, R.: Domain-specific knowledge acquisition and classification using wordnet. In: Etheredge, J.N., Manaris, B.Z. (eds.) Proceedings of the Thirteenth International Florida Artificial Intelligence Research Society Conference, May 22-24, 2000, Orlando, Florida, USA. pp. 224–228. AAAI Press (2000), http://www.aaai.org/Library/FLAIRS/2000/flairs00-043.php
15. Noy, N.F., McGuinness, D.L., et al.: Ontology development 101: A guide to creating your first ontology (2001)
16. Qawasmeh, O., Lefrançois, M., Zimmermann, A., Maret, P.: Computer-assisted ontology construction system: Focus on bootstrapping capabilities. In: The Semantic Web: ESWC 2018 Satellite Events - ESWC 2018 Satellite Events, Heraklion, Crete, Greece, June 3-7, 2018, Revised Selected Papers. pp. 60–65 (2018). https://doi.org/10.1007/978-3-319-98192-5_12, https://doi.org/10.1007/978-3-319-98192-5_12
17. Vandenbussche, P., Atemezing, G., Poveda-Villalón, M., Vatant, B.: Linked open vocabularies (LOV): A gateway to reusable semantic vocabularies on the web. Semantic Web **8**(3), 437–452 (2017). https://doi.org/10.3233/SW-160213, https://doi.org/10.3233/SW-160213

18. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014). https://doi.org/10.1145/2629489, http://doi.acm.org/10.1145/2629489