Computer-assisted Ontology Construction System: Focus on Bootstrapping Capabilities

Omar Qawasmeh[1], Maxime Lefrançois[2], Antoine Zimmermann[2], Pierre Maret[1]

[1] Univ. Lyon, CNRS, Lab. Hubert Curien UMR 5516, F-42023 Saint-Étienne, France
{omar.alqawasmeh,pierre.maret}@univ-st-etienne.fr
[2] Mines Saint-Etienne, Univ. Lyon, Univ. Jean Monnet, IOGS, CNRS, UMR 5516, LHC, Institute Henri Fayol, F-42023 Saint-Étienne France
{maxime.lefrancois,antoine.zimmermann}@emse.fr

Abstract. In this research, 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 large public knowledge bases. We report on a comparative study between our system and the existing ones on the wine ontology.

Keywords: Ontology construction, Knowledge base

1 Introduction

Ontologies play nowadays an important role in organizing and categorizing data in information systems and on the web. This leads to a better understanding, sharing and analyzing of knowledge in a specific domain. As mentioned in [1], 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. The process of ontology development is facing two main problems: the initiation of the extraction phase (cold start, blank page problem) [2], and the large number of micro-contributions that the domain experts must do. These problem are addressed by automatic or semi-automatic ontology development systems, that help in avoiding the cold start, and in minimizing the time spent by the domain experts. In this paper we propose the design of a new functionality focusing on the bootstrapping and combined with interactions with the knowledge engineer. Our functionality takes advantage of three large public knowledge bases: a) DBpedia [3], b) Wikidata [4] and c) NELL (Never Ending Language Learner) [5]. We report on the evaluation of our functionality compared with other approaches, using the ontology for wine. The rest of this paper is organized as follows: Section 2 presents a short-state of the art in the field, Section 3 depicts our designed system, Section 4 reports on the results of experiments for evaluation, and Section 5 concludes the paper.

2 Automatic Ontology Development: A State of the Art

Bedini et al. [6] define four categories to classify the approaches for automatic ontology development: 1. Conversion or translation, 2. Mining based, 3. External knowledge based,

and 4. Frameworks. We shortly present here a set of approaches that are related to our approach technique (External knowledge based). Kong et al. [7] use WordNet [8] 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 1 shows their results comparing to the wine ontology. Kietz et al. [9] propose an approach that uses three knowledge bases to construct ontologies. They used a generic ontology to generate the main structure, a dictionary containing generic terms close to the required domain, and 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). Cahyani and Wasito [10] propose an automatic system to build an ontology for the Alzheimer's disease. Their system consists of the following steps: 1. a term relation extraction to match the extracted relations to Alzheimer glossary [4]. 2. matching with ontology design patterns. 3. builds and evaluate the ontology. 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. We propose in the next section an original functionality for semi-automatic ontology development tools.

3 A semi-automatic Approach for Bootstrapping Ontology

As shown from the literature review, most of the approaches considering external knowledge bases 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). In order to improve current automatic ontology construction, we propose a functionality using publicly available knowledge bases: DBpedia, Wikidata and NELL [5]. The pros of using these knowledge bases are that they are structured, very large, include rich relations, evolving in time, machine understandable and multilingual.

We follow a semi-automatic bootstrapping technique, where the user enters a set of keywords related to a specific domain (e.g. wine, grapes, and wine color, for the wine domain). Then by issuing a series of queries to the external knowledge bases, several classes and relations are extracted. Then the generated list is shown to the user for selection(see Figure 1). After that, the set of classes is used to extract the instances from the NELL knowledge base. Our process is described in Algorithm 1. In the following subsections we present different phases implemented.

3.1 Extract General Information (DBpedia)

DBpedia knowledge base [3] contains structured information from Wikipedia that is accessible via a SPARQL endpoint[11]. In this phase, the set of keywords are used to perform queries over the DBpedia knowledge base to get some information that will help the user to choose clearly among the related terms that can be retrieved. For

[3] https://www.w3.org/TR/owl-guide/wine.rdf
[4] https://www.alz.org/care/alzheimers-dementia-glossary.asp
[5] An executable jar file of our algorithm can be found here https://goo.gl/vCj3rU

Algorithm 1: The General Algorithm Implemented by our System

```
ConstructInitialOntology(keywords);
Input  : keywords, a list of keywords given by the domain expert
Output : ⟨classes,relations,instances⟩ lists of terms.
⟨classes,relations,instances⟩ ← ⟨∅,∅,∅⟩
foreach keyword in keywords do
    ⟨abstract,labels,uri⟩ ← queryDBPedia(keyword);                    // see section 3.1
    ⟨classes,relations⟩ ← queryWikiData(keyword);                     // see section 3.2
    instances ← queryNELL(keyword) ;                                  // see section 3.3
    ⟨classes′,relations′,instances′⟩ ← pick(abstract,labels,uri,classes,relations,instances);
     // user picks
    classes ← classes∪classes′;
    relations ← relations∪relations′;
    instances ← instances∪instances′;
return ⟨classes,relations,instances⟩ ;
```

example, the output for the keyword "wine" is: the abstract from wine's Wikipedia page [6], the label in DBpedia in any supported language, and the different types from DBpedia (e.g. beverage, food).

3.2 Extract Classes and Relations (Wikidata)

Wikidata [4] is a collaborative, multilingual, structured knowledge base that can be read and modified by both humans and machines. The information on Wikidata is accessible by querying services. An initial query to Wikidata returns us the IDs of the users' keywords. Then, using these IDs, we perform different queries over the Wikidata to retrieve a set of classes and the relations. We use three different queries to have the following output: 1. Classes, with the parent-child relationship. For instance, the query was able to retrieve 80 different classes for the keyword "wine". 2. The most connected relations for each class. A list of relations that are connected to a specific class is retrieved along with the number of instances that are using this relation. For instance, the query with "wine" retrieves 6 different relations and their number of use. 3. Classes, along with their top-level high classes. A list of relations that are connected to two different classes are retrieved along with the number of instances that are using this relation. For example for the class wine and the class alcoholic beverage the query was able to retrieve 7 different subclasses.

3.3 Extract Instances (NELL)

Since January 2010, a computer system called NELL (Never-Ending Language Learner) [5] has been running continuously, in order to learn over time from the World Wide Web. NELL currently has more than 50 millions beliefs [7], which are attached to different levels of confidence, and features. We use three main files to access NELL: 1. Relations: contains 460 relations that were extracted manually. 2. Categories: contains 291 categories that were extracted manually. 3. Instances: contains 2,971,069 instances. In this phase, we use the NELL knowledge base in order to build a candidate list of

[6] https://en.wikipedia.org/wiki/Wine Last visit Jan-2018
[7] Based on: http://rtw.ml.cmu.edu/rtw/ Last visit: Oct-2017

instances that are related to the given set of keywords. NELL is queried based on a set of features such as domain, range, and confidence values. The next section discusses the initial experiments we use to validate our functionality.

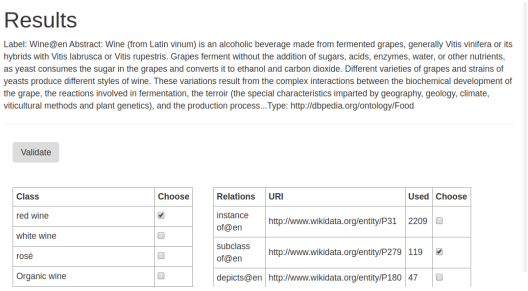

Fig. 1: A subset of the classes and relations that are extracted for the keyword wine.

4 Evaluation and Demonstration

In order to validate our approach, we compare our results to those published in [7](See section 2). We therefore lead a similar experiment to evaluate our system, and we compare our results to the baseline ontology [8] and to the results in [7]. Authors in [7] use keyword "wine" to perform a query over WordNet. So that the comparison is 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[9]. Table 1 gives an overview of these results are compare them to the W3C's wine ontology and to the results of [7]. 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 to filter NELL's beliefs. This experiment shows that our system performs better than [7] while proposing only relevant concepts, which allows us to assert it would be a good fit for the bootstrapping phase of ontology development. As for the demonstration experiments, a set of tasks could be done such as: let the users to choose a specific domain to test the functionality of the system, or to regenerate the experiments we already did on the wine domain.

5 Conclusion and Future work

In this paper we propose an original approach for ontology bootstrapping based on the use of three external knowledge bases: DBpedia, WikiData, an NELL. Preliminary results shows that our system performs better than [7] that is based on WordNet. 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.

[8] https://www.w3.org/TR/owl-guide/wine.rdf
[9] "wine" experiment: full lists of terms our System outputs http://bit.ly/2EEKItn

Approach	W3C's wine ontology	[7]'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, [7]'s approach and the W3C's wine ontology

Relation	**Count**	**URI of the relation**
instance of	2254	http://www.wikidata.org/entity/P31
subclass of	96	http://www.wikidata.org/entity/P279
depicts	35	http://www.wikidata.org/entity/P180
main subject	8	http://www.wikidata.org/entity/P921
has part	6	http://www.wikidata.org/entity/P527
material used	6	http://www.wikidata.org/entity/P186

Table 2: Set of RDF-Relations Extracted for the keyword wine

As for future work, we plan to extend the number of external knowledge bases that we query, to support the collaborative functionalities between the different parties, and to provide a web service for the functionality.

References

1. E. Blomqvist, "Pattern ranking for semi-automatic ontology construction," in *Proceedings of the 2008 ACM Symposium on Applied Computing, Brazil*, 2008.
2. Y. Zhang, T. Tudorache, M. Horridge, and M. A. Musen, "Helping users bootstrap ontologies: An empirical investigation," in *the 33rd Annual ACM Conf. on Human Factors in Computing Systems, Seoul, Republic of Korea*, 2015.
3. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives, "Dbpedia: A nucleus for a web of open data," in *The Semantic Web, 6th Int. Semantic Web Conf., 2nd Asian Semantic Web Conf., ISWC 2007 + ASWC 2007, Busan, Korea,*, 2007.
4. D. Vrandecic and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, 2014.
5. A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell, "Toward an architecture for never-ending language learning," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, USA*, 2010.
6. I. Bedini and B. Nguyen, "Automatic ontology generation: State of the art," *PRiSM Laboratory Technical Report. University of Versailles*, 2007.
7. H. Kong, M. Hwang, and P. Kim, "Design of the automatic ontology building system about the specific domain knowledge," in *Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference*, IEEE, 2006.
8. G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
9. J.-U. Kietz, A. Maedche, and R. Volz, "A method for semi-automatic ontology acquisition from a corporate intranet," in *EKAW-2000 Workshop "Ontologies and Text", Juan-Les-Pins, France*, 2000.
10. D. E. Cahyani and I. Wasito, "Automatic ontology construction using text corpora and ontology design patterns (odps) in alzheimer's disease," *Jurnal Ilmu Komputer dan Informasi*, 2017.
11. S. Harris, A. Seaborne, and E. Prud'hommeaux, "Sparql 1.1 query language," *W3C recommendation*, vol. 21, no. 10, 2013.