A Framework for Ontology Learning from Taxonomic Data

Chandan Kumar Deb, Sudeep Marwaha, Alka Arora and Madhurima Das

Abstract Taxonomy is implemented in myriad areas of biological research and though structured it deals with the problem of information retrieval. Ontology is a very powerful tool for knowledge representation and literature also cites the conversion of taxonomies into ontologies. The automated ontology learning is developed to ward off the knowledge acquisition bottleneck; but thereof the limitation includes text understanding, knowledge extraction, structured labelling and filtering. The system, ASIUM, TEXT TO ONTO, DODDLE II, SYNDIKATE, HASTI, etc., includes some inadequacies and does not exclusively deal with taxonomic texts. The proposed system will deal with the taxonomic text available in agricultural system and will also enhance the algorithms thereby available. We also propose a framework for learning of the taxonomic text which will overcome the loopholes of ontology developed from generalized texts. Finally, a framework of comparison of the manually developed ontology and automatically developed ontology will be ensured.

Keywords Automated ontology learning · Taxonomic texts · Knowledge acquisition

C.K. Deb (✉) · S. Marwaha · A. Arora
Indian Agricultural Statistics Research Institute, New Delhi, India
e-mail: chandan@iasri.res.in

S. Marwaha
e-mail: Sudeep.Marwaha@icar.gov.in

A. Arora
e-mail: Alka.Arora@icar.gov.in

M. Das
Indian Agricultural Research Institute, New Delhi, India
e-mail: madhurima.iari@gmail.com

V.B. Aggarwal et al. (eds.), *Big Data Analytics*, Advances in Intelligent
Systems and Computing 654, https://doi.org/10.1007/978-981-10-6620-7_4

1 Introduction

Nowadays, knowledge structuring and knowledge management are the key focuses of the scientific communities. Ontology is a very powerful knowledge representation technique. On the other hand, the taxonomic knowledge has a great correspondence to the ontology. As [1] proposed a methodology for the conversion of taxonomies into ontologies; but manual ontology building is a tremendous labour intensive task. Although unstructured data can be made into structured; it encompasses a very lengthy process and henceforth the automated ontology learning approach is developed to ward off this knowledge acquisition bottleneck, hitherto, it includes some serious limitation in text understanding, knowledge extraction, structured labelling, and filtering [2]. Under conventional condition, the ontology learning deals with the normal text. Ontology learning from normal text is not so efficient. It is also dangerous to extract the concept from the normal text. However, no attempt has been made for ontology learning from taxonomic text. Thus, a novel approach is proposed to engineer the taxonomic text and make it as the input of the ontology learning. In agriculture, this kind of ontology learning has not yet been attempted.

2 Literature Review

Ontology learning is a new field of artificial intelligence and machine learning. A limited number of ontology learning tools and techniques have been developed so far and some of them are listed below:

References [3, 4] developed a system namely ASIUM. ASIUM learns sub-categorization frames of verbs and ontologies from syntactic parsing of technical texts in natural language. It is developed in French language. The ASIUM method is based on conceptual clustering. Reference [5] developed a system to classify nouns in context. It is able to learn categories of nouns from texts, whatever their domain is. Words are learned considering the contextual use of them to avoid mixing their meanings. This system was a preprocessor of ontology learning. References [6, 7] developed a system of ontology learning named TEXT TO ONTO. TEXT TO ONTO learns concepts and relations from unstructured, semi-structured, and structured data, using a multi-strategy method which is a combination of association rules, formal concept analysis, and clustering. But this is based on the shallow natural language processing. This system fails to address complex levels of understanding. Mostly, it identified concepts through regular expression. Reference [8] developed a system namely DODDLE II. DODDLE II is a Domain Ontology Rapid Development Environment. It can construct the hierarchical and nonhierarchical relationship of the domain concepts. For the hierarchical relationship, it uses WordNet. References [9–11] developed a system namely SYNDIKATE. SYNDIKATE is a system for automatically acquiring knowledge from real-world texts. It is available in German language. It has the problem of co-reference resolution. References [12, 13] developed a system namely HASTI. HASTI is an automatic ontology building system, which builds dynamic ontologies from scratch. HASTI learns the lexical and ontological knowledge from natural language texts. This is available in the Persian language.

Reference [14] developed a system that integrates machine learning and text mining algorithms into an efficient user interface; lowering the entry barrier for users who are not professional ontology engineers. The main features of the systems include unsupervised and supervised methods for concept suggestion and concept naming, as well as ontology and concept visualization. Reference [15] developed a system that integrates the external source knowledge like DBPedia and OpenCyc for getting the automatic suggestions for labelling the concepts. Reference [16] discussed how to learn large-scale ontology from Japanese Wikipedia. The large ontology includes IS-A relationship; class–instance relationship; synonym; object and data type properties of domain. However, a big problem of weakness in upper ontology arose against building up higher quality general ontology from Wikipedia. Reference [17] proposed a novel model of an *Ontology Learning Knowledge Support System (OLeKSS)* to keep the Knowledge Support System updated. The proposal applies concepts and methodologies of system modelling as well as a wide selection of ontology learning processes from heterogeneous knowledge sources (ontologies, texts, and databases), in order to improve KSS's semantic product through a process of periodic knowledge updating.

Reference [18] developed a semi-supervised ontology learning based focused (SOF) crawler. This embodies a series of schemas for ontology generation and web information formatting. In this system, the web pages are segregated by Support Vector Machine (SVM). Reference [19] proposed a ontology learning approach that has been used for developing the ontology. They used Linking Open Data (LOD) cloud which is a collection of Resource Description Framework (RDF). They used domain ontology for learning ontology and called Mid-Ontology Learning. Mid-Ontology learning approach that can automatically construct a simple ontology, linking related ontology predicates (class or property) in different data sets. Reference [20] gave an approach of clustering of the web services for efficient clustering. They adopted the ontology learning to generate ontologies via hidden semantic pattern. But they also mentioned the chances of failure of the ontology based discovery of web services. Reference [21] used heterogeneous sources like databases, ontologies, and plain text for ontology learning. Reference [22] generated ontology structure called ontology graph. The ontology graph defines ontology and knowledge conceptualization. The ontology learning process defines the method of semiautomatic learning and generates ontology graphs from Chinese text of different domains.

3 Objectives of the Proposed Work

The proposed system will deal with the taxonomic text available in agricultural system. We also propose a framework for learning of the taxonomic text which will overcome the loopholes of ontology developed from generalized texts. One system has been developed on the basis of the learning frame work. Finally, a framework of comparison of the manually developed ontology and automatically developed ontology will be ensured.

4 Proposed Framework

This proposed framework will mainly differ from the conventional ontology learning process in the input of the ontology learning framework. The conventional ontology learning system claims to have the capability of dealing with a range of texts but these systems are trapped by the inherent hindrance of the ontology learning. On the other hand, this framework is totally focused on the taxonomic text available in agricultural system. The proposed framework will mainly deal with two sub-module—first, how to deal with the taxonomic text and second, how to validate of the result of the first module.

4.1 Algorithmic Framework

This proposed framework subdivides the ontology learning process into well-demarcated category. It neutralizes the complexity of the ontology learning process. Figure 1 shows the schematic diagram of the total framework.

4.1.1 Categorize the Taxonomic Data

The taxonomic texts are available in different forms or categories. The first task of this framework is to find out the category of the taxonomic text on the basis of different sources (e.g. Taxonomic Books in Agriculture). Based on this category, the whole process of the ontology learning will be done.

4.1.2 Preprocess the Text

Preprocessing of the taxonomic data is very important because whatsoever the source of the data and category they have; there exists two basic type of text—

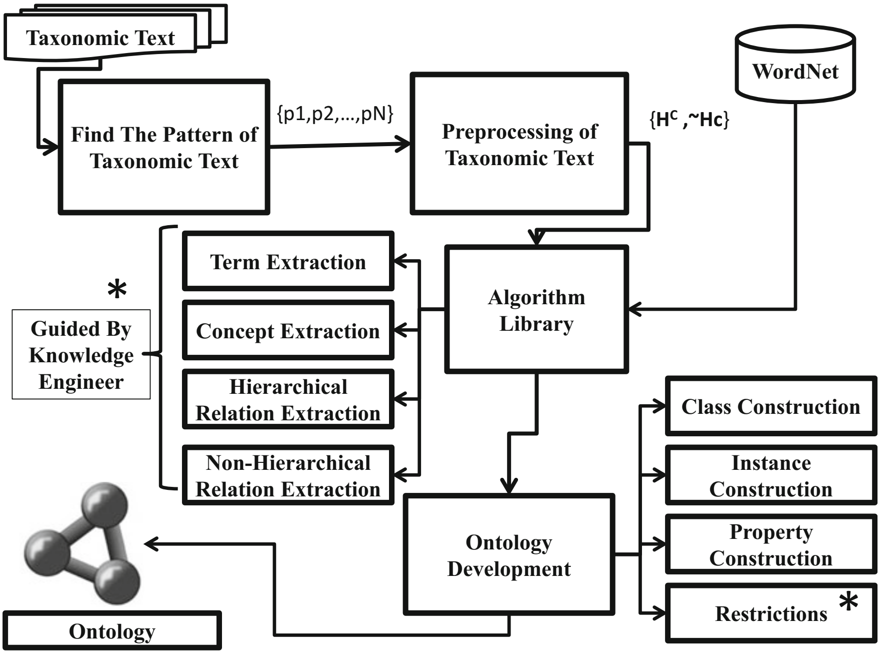

Fig. 1 Schematic task flow of algorithmic framework

hierarchical and nonhierarchical. For this, the ontology engineer can use any algorithm which can be helpful for the partitioning of the data (e.g. SVM).

4.1.3 Development of the Algorithmic Library

First part of the total framework, i.e. Algorithmic Framework wholly deals with the algorithms for the taxonomic text ontology learning; it deals with the following task.

Term Extraction

This sub-module actually commences the ontology learning process; it is from this level that the extraction of the ontology building block is started. The term extracted is used for class and instance construction. A repository will be developed for the extracted term. For extracting the term, the tools and techniques of natural language processing can be used.

Concept Extraction

Next step towards ontology learning is the extraction of the concept. The concept extraction can be done in two ways—the first approach is the use of the taxonomy for the concept labelled and second with the help of WordNet API like JNWL.

4.1.4 Relationship Extraction

In preprocess module, the text is subdivided into two categories:

Hierarchical Relation Extraction

On the basis of the pattern of the data, the hierarchical or ISA relation will be extracted. These algorithms will be based on the taxonomic data so these relations will also be extracted from the basis of taxonomic data.

Nonhierarchical Relation Extraction

Apart from the hierarchical relation or ISA relation, there are many relations like hasA, partOf for construction of the ontology.

4.1.5 Mapping to Ontology

After the extraction process of term and concept; the class and subclasses will be constructed for developing the ontology. The identification of the properties is also a subtask of this task. Restrictions will be imposed on the class by the help of the knowledge engineer.

4.2 Architecture of Proposed System

The given framework has been implemented in MVC architecture. Here, we proposed java-based *N*-tier architecture. Different layer has its individual importance as well as they are important as a whole. The layers are as follows and Fig. 2 depicts the architecture of the software:

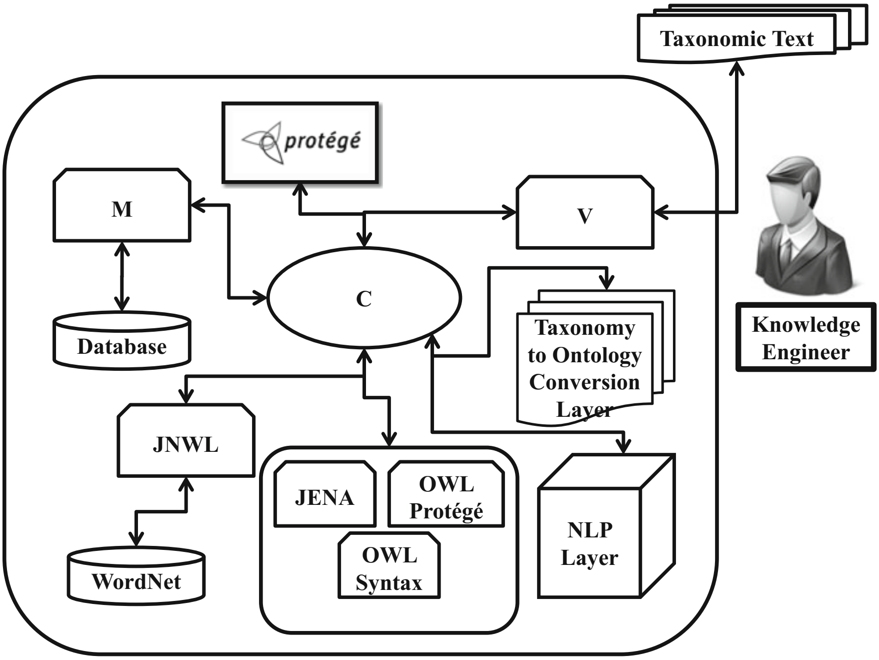

Fig. 2 Architecture of the system

4.2.1 MVC Layer

In this layer, the model, view, and controller are developed. The controller part of MVC interacts with the other part of the software and it works as a central control to the whole system.

4.2.2 Natural Language Processing Layer

The layer of this architecture is important, because all the preprocessing before automatic ontology development, it extracts all the building block of ontology. The term extraction, concept extraction, etc., are given in the framework and have been implemented in this layer.

4.2.3 Tax-to-Onto Layer and Semantic Web Layer

This layer deals with some very important tasks and components of ontology learning. It is connected with protégé which is the knowledge base of the ontology. It has several API's as its components (e.g. JENA, OWLProtege, JNWL and Wordnet).

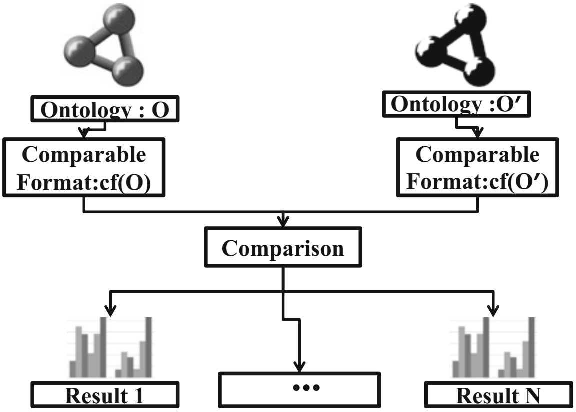

Fig. 3 Comparisons of the ontologies

4.3 *Comparison Framework*

This framework provides the approach of comparison between the two ontologies. By this approach, we can also evaluate our framework of ontology learning from taxonomic text. For comparing the ontology, we have to convert both the ontology into a single comparable format. The comparable format may be the conversion of ontology to a graph. Then compare both the ontology on the basis of class, instance, and properties. Comparison can also be done on the basis of concept extraction. Figure 3 depicts the comparison framework of ontology.

5 Conclusion

The software developed on the basis of the proposed framework will help in automatic ontology learning from taxonomic texts and also overcome the inherent problems of conventional ontology learning in terms of knowledge acquisition. The proposed methodology may be used in the biological fields, where taxonomy has its own importance. Besides biological field, this methodology is generic enough to be applied in other fields also. Lastly, this framework also provides a more simplistic way of comparison between manually developed and automatically developed ontology.

Acknowledgements The first author gratefully acknowledges the INSPIRE Fellowship provided by Department of Science and Technology, New Delhi.

References

1. Bedi, P., Marwaha, S.: Designing ontologies from traditional taxonomies. In: Proceedings of International Conference on Cognitive Science, Allahabad, India (2004)
2. Zouaq, A., Gasevic, D., Hatala, M.: Unresolved issues in ontology learning. In: Proceedings of Canadian Semantic Web Symposium (2011)
3. Faure, D., Nédellec, C., Rouveirol, C.: Acquisition of semantic knowledge using machine learning methods: The system "ASIUM". In Universite Paris Sud (1998)
4. Faure, D., Thierry, P.: First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. In: Ontology Learning ECAI-2000 Workshop (2000)
5. Chalendar, G., Grau, B.: Knowledge engineering and knowledge management. In: Methods, Models and Tools. Springer, Berlin (2000)
6. Maedche, A., Staab, S.: Discovering conceptual relations from text. In: Proceedings of the 13th European Conference on Artificial Intelligence. IOS Press, Amsterdam (2000)
7. Maedche, A., Volz, R.: The ontology extraction maintenance framework text-to-onto. In: Proceedings of the Workshop on Integrating Data Mining and Knowledge Management (2001)
8. Yamaguchi, T., Izumi, N., Fukuta, N., Sugiura, N., Shigeta, Y., Morita, T.: DOODLE-OWL: OWL-based semi-automatic ontology development environment. In: Proceedings of the 3rd International Workshop on Evaluation of Ontology based Tools (2001)
9. Hahn, U., Schnattinger, K.: Towards text knowledge engineering. In: Proceedings of the 15th National Conference on Artificial Intelligence, Madison, Wisconsin (1998)
10. Hahn, U., Romacker, M.: The SYNDIKATE text Knowledge base generator. In: Proceedings of the 1st International Conference on Human Language Technology Research (2001)
11. Hahn, U., Markó, K.: An integrated dual learner for grammars and ontologies. Data Knowl. Eng. **42**, 273–291 (2002)
12. Mehrnoush, S., Ahmad, B.: An introduction to hasti: an ontology learning system. In: Proceedings of the Iasted International Conference Artificial Intelligence and Soft Computing. Acta Press, Galgary, Canada (2002)
13. Mehrnoush, S., Ahmad, B.: The state of the art in ontology learning: a framework for comparison. Knowl Eng Rev **18**, 293–316 (2003)
14. Fortuna, B., Grobelnik, M., Mladenić, D.: OntoGen: Semi-automatic ontology editor. In: Proceedings of Human Interface, Part II, HCI International, LNCS 4558 (2007)
15. Weichselbraun, A., Wohlgenannt, G., Scharl, A.: Refining non-taxonomic relation labels with external structured data to support ontology learning. Data Knowl. Eng. **69**, 763–778 (2010)
16. Tamagawa, S., Sakurai, S., Tejima, T., Morita, T., Izumiy, N., Yamaguchi, T.: Learning a large scale of ontology from Japanese Wikipedia. In: Proceedings of International Conference on Web Intelligence and Intelligent Agent Technology (2010)
17. Gil, R.J., Martin-Bautista, M.J.: A novel integrated knowledge support system based on ontology learning: model specification and a case study. Knowl. Based Syst. **36**, 340–352 (2012)
18. Dong, H., Hussain, F.K.: SOF: a semi supervised ontology earning based focused crawler. Concurr Comput Pract Experience **25**, 1755–1770 (2013)
19. Lihua, Z.H.A.O., Ichise, R.: Integrating ontologies using ontology learning approach. IEICE Trans Inf Syst **96**, 40–50 (2013)
20. Kumara, B.T., Paik, I., Chen, W., Ryu, K.H.: Web service clustering using a hybrid term-similarity measure with ontology learning. Int. J. Web Serv. Res. **11**, 24–45 (2014)
21. Gil, R., Martin-Bautista, M.J.: SMOL: a systemic methodology for ontology learning from heterogeneous sources. J Intel Inf Syst **42**, 415–455 (2014)
22. Liu, J.N., He, Y.L., Lim, E.H., Wang, X.Z.: Domain ontology graph model and its application in Chinese text classification. Neural Comput. Appl. **24**, 779–798 (2014)