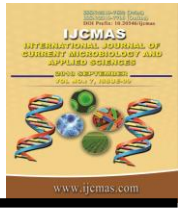
Int.J.Curr.Microbiol.App.Sci (2018) *7*(9): 975-985

International Journal of Current Microbiology and Applied Sciences
***ISSN: 2319-7706* Volume 7 Number 09 (2018)**
Journal homepage: http://www.ijcmas.com



Original Research Article https://doi.org/10.20546/ijcmas.2018.709.117

Building Soil Taxonomy Ontology by the Way of Connective Based Ontology Learning

Chandan Kumar Deb[1*], Madhurima Das[2] and Sudeep Marwaha[1]

[1]Indian Agricultural Statistics Research Institute, New Delhi-110012, India
[2]Indian Agricultural Research Institute, New Delhi-110012, India

**Corresponding author*

Keywords

Ontology, Ontology learning, Connectives, Soil ontology

Article Info

Accepted:
08 August 2018
Available Online:
10 September 2018

A B S T R A C T

Ontology is one of the most popular knowledge representation techniques. Ontology based knowledgebase system is very well structured. Building of knowledgebase manually is a quite cumbersome process. This knowledge acquisition bottle neck can be overcome by making the ontology building process automated. Accordingly, the ontology learning came into the scene. In ontology learning, the natural text is taken as an input and building of the ontology is done. One of its drawbacks is that taking input from natural text to ontology development several limitations are being encountered. In this research work, we have taken two aspects of ontology building. We have firstly inducted taxonomy and secondly extracted the property of the taxonomy automatically from the semi structured text. For demonstration purpose, we have taken USDA soil taxonomy; in which we run our algorithm in a single chapter 'Alfisol' and got a very significant result. This novel method of ontology building process based on connectives facilitate us the minimal use of corpus thus waiving of its tediousness.

Introduction

Drawing a minimal reasoning by an intelligent organism required to use a massive background knowledge that will act as a support for proper reasoning. This is an era of developing intelligent system that act as a human helping in dealing of huge data which is singly impossible to manage by a human being. In this AI (Artificial Intelligence) era, the machine learning concept became evident in the scenario. Ontology learning is a part of that journey. Automated ontology learning became an unavoidable part for creating knowledgebase efficiently and expeditiously. Ontology learning involves several detailed subtasks. It includes taxonomy induction and property identification which are inevitable tasks for making an ontology learning system.

Previously several attempts have been made for building taxonomy and property extraction from natural text. The taxonomy induction from the natural text revolves around a broad range of approaches. Semantic approach is mainly focused to induct taxonomy automatically to construct the knowledge repository. Ponzetto *et al.*, (2011) constructed taxonomy from Wikipedia and compared their taxonomy with manually developed lexical

ontology i.e. Word Net. Another semantic approach taxonomy induction tool Taxo Learn; developed by Dietz *et al.,* (2012) used hierarchical clustering to induct the taxonomy from text. Medelyan *et al.,* (2013) created a focused taxonomy using Natural Language Processing Tool. A series of work such as Rios-Alvarado *et al.,* (2013); Liu *et al.,* (2013); Meijer *et al.,* (2014); and Hosny *et al.,* (2015) inducted taxonomy using semantic approach from natural language which are either semiautomatic or automatic. In 2011 some of the authors used probabilistic model to develop taxonomy from the text Fallucchi *et al.,* (2011). Another very important approach of the taxonomy induction is the graph based approach. Using hierarchical random graphs Fountain *et al.,* (2012) developed a system that is capable of induction of automatic lexical taxonomy from natural text. Onto Learn is a graph based system developed by Velardi *et al.,* (2013) which can find the taxonomy from the scratch. Taxo Finder is another taxonomy learning system. This is also a graph based system which builds a graph and measures the association of the concepts. Novelli *et al.,* (2012) developed a system that can extract the concept, class, properties from the natural text. Literatures depict the tediousness and time consuming nature of ontology building from text. To overcome this knowledge acquisition bottleneck several scientific communities have built many systems that help in the ontology building without or little intervention of the human being. The ontology building frame work takes a huge resource and labour to make the system efficient. The key task of ontology building is corpus based which is very much general in nature. If some research community try to work with the more focused field it is very difficult to identify the taxonomy. But the avoidance of the use of the corpus is almost impossible in the ontology learning where natural language processing is the key task.

In this research work we induct the taxonomy from the taxonomic text as suggested by the Deb *et al., (*2015) from conversion of Taxonomies into Ontologies but in automated manner.

This research work mainly deals with two objectives: firstly, the induction of taxonomy from taxonomic text and secondly the extraction its properties. We have inducted taxonomy from the taxonomic text and properties by using connectives and thus the mundane work of corpus development has been easily waived of.

Materials and Methods

Taxonomic text

From our previous observation Deb *et al.,* (2015) we have described that the property of the taxonomic text may be exploited for ontology learning. In this research work we have selected USDA soil taxonomy book for the input of the text of the ontology learning framework.

Natural language or unstructured natural language is very difficult to be parsed because of its ambiguous nature. We have chosen taxonomic text which is a semi structured language source for automated ontology development.

Architecture of the developed software

The main component of the architecture of the system is the involvement of OpenNLP to the basic java compiler (Fig. 1). This helps in doing the NLP task as per the requirement of the system. Another highlight of the system is the algorithm library which implicitly contains the NLP unit and this library is plugged in to the system. Other components are the API's like JENA, OWL Protégé and OWL Syntax are use to deal with the protégé and ontology.

We have developed a system that is enabled of extracting taxonomic class and property.

Schematic process flow of the software development

The process of the software development involves the following described tasks (Fig. 2).

It includes the following steps.

Step 1a and Step 1b: Segregation

This is the first step of ontology learning process. In this step total text or the part of the text which is under processing is divided into two parts i.e. the text containing the taxonomy and the text without the taxonomy. The text is segregated on the basis of connectives present in the sentence. A connective is a set that contains the key words that help to identify the relationship among the class.

Step 2a and Step 2b: Sentence detection

Sentence detection is the next step of the ontology learning algorithm. Segmentation of the total text into sentences is done for further task of NLP.

Step 3a: Tokenization

The sentence is further subdivided into words and single symbol called tokenization.

Step 4a: Parts-of-speech tagging

In this task of NLP we find the proper noun for the identification of the taxonomic class of the taxonomic text.

Step 5a: Name entity recognition

After Step 4a Name Entity Recognition is very important because name extraction critically depends on domain. For detection of name the corpus can be built on the basis of the corresponding domain.

Step 6a: Hierarchical class recognition

In this step, it is studied based on the relationship or the parent child relationship of the extracted name.

Open NLP

Natural Language Processing (NLP) is very important in the field of Artificial Intelligence. In the human computer interaction NLP is a major growing field. Through NLP, human can achieve many of the Interaction Bridge between human and machine. Important information can be extracted from the Natural Text and it can be used for mankind through NLP. From implementation point of view there are many tools available in web for implementing the Natural Language Processing. Like other tools Open NLP is a tool for natural language processing. This tool was released by Apache Foundation. First stable release was done in April, 2013. Open NLP is a Java based tool. It is an open source; jar is freely available and it easily integrates to the application. Open NLP is able to do most of NLP task like segmentation, tokenization, part-of-speech tagging, named entity extraction, chunking, parsing, and co-reference resolution. All the above mentioned tasks are done by the help of previously developed models. Particular task has a particular predefined model. User also can develop their model for a particular NLP task.

Connectives

In previous section we have discussed the importance of the taxonomic text for building domain ontology. We have observed that the taxonomy available in the literature follows some definite pattern. Connectives are the

integral part of that pattern.
We have observed that the connectives are uniformly distributed over text and are used for a particular task in the text. Following paragraph describe about connectives and the importance of this.

Connectives are the special kind of word or group of words that connects objects to describe a particular type of relationship in a particular domain.

$^{T}C_{D}$: Connectives (i)

Hierarchical Connectives

These are those connectives that describe the hierarchical or the parent child relationship from the text.

$^{H}C_{ST}$: This is the superscript H described Hierarchical Connectives

Property Based Connectives

These are those connectives that describe the property of the taxonomic class.

USDA Soil Taxonomy

Aqualfs are the Alfisols that have aquic conditions for some time in normal years (or artificial drainage) at or near the soil surface.

Albaqualf are the Aqualfs with ground water seasonally perched above a slowly permeable argillic horizon

The above text snippet shows that the “are the” and “that have” helps to identify the parent child relationship of the taxonomy and their property. Fortunately this pattern is present throughout the text. This is a special characteristic of taxonomic text (Deb *et al.,* 2015). The “are the” and “that have” one of the example of hierarchical and property based connectives.

Results and Discussion

This research works mainly deals with two parts - first task is the induction of the taxonomy and second task is to extract property from the taxonomic text. Both the above described task involves the following activities.

Database design of the software

The Figure 3 describes the ER-diagram of the system that can capture the necessary information in a structured manner to accelerate the work flow of the developed algorithms. The database captures the sentences which are detected by the algorithms. It captures the tokens of the tokenized sentence. It helps to isolate the classified (Hierarchical and Non-hierarchical) and also helps to map the token to the exact parts of speech and connectives with the help of the developed algorithms. It can also capture the property of the classes. Finally the database helps to identify the ontology classes and properties of the de-tokenized sentences for finding the hierarchical and property pattern.

Class diagram of the software

The class is the core part of our developed system. It consists of several classes for implementation of our algorithm. For implementing the class we used JDK 8.0 and plug the Open NLP which has already been described previous section. Class diagram do the entire task which will be discussed in the algorithm section (Fig. 4).

Taxonomy Induction

The important step of the algorithm is the pre-processing of natural language. Input to any natural language processing has to be pre-processed. The normalized text or the pre-

processed text are free from typo mistakes, unnecessary data like page number, diagram and other noise which is not the part of the natural language. Another important process of the algorithm is - part of speech tagging and connectives tagging. In this step we try to tag the POS and domain based connectives. After this step we again de tokenized the sentence with parts of speech and the connectives. Then we have to find the particular pattern that depicts the parent child relationship. Thereafter we have to identify the parent child and the connectives. Create the three node binary tree where the root node is the connective node. Create all the binary tree and feed into the further steps to super impose the binary and create a large graph to depict the whole taxonomy. If the nodes are not matched then create a new tree for the independent tree.

In this research work we have only focused on the extraction of the hierarchical relationship from among the objects. We have conducted our experiment on USDA Soil Taxonomy and took a single chapter namely "Alfisol" to identify the hierarchy available in that chapter.

We apply our algorithm to the chapter and segregate 3342 sentences from the chapter. Algorithm suggests that among 3342 sentences 626 sentences indicate the hierarchical or parent child relationship on the basis of keywords (Fig. 5).

Initially the algorithm take the sentences as an input and roughly classify the text into two broad group i.e. the hierarchical and the non-hierarchical. After that the entire hierarchical sentence are traversed one by one and creation of the three node binary tree is done. Figure 6 describes the creation of a binary tree where <HC> is the root node and the left child of the tree is Aqualf and right child of the tree is Alfisol. Basically this binary tree has a rule to identify the taxonomic parent child relationship. The left child of the tree is the child of the right child of the tree. Likewise all the trees has been created by the algorithm.

Fig.1 Schematic representation of N - tier architecture of the software using Open NLP

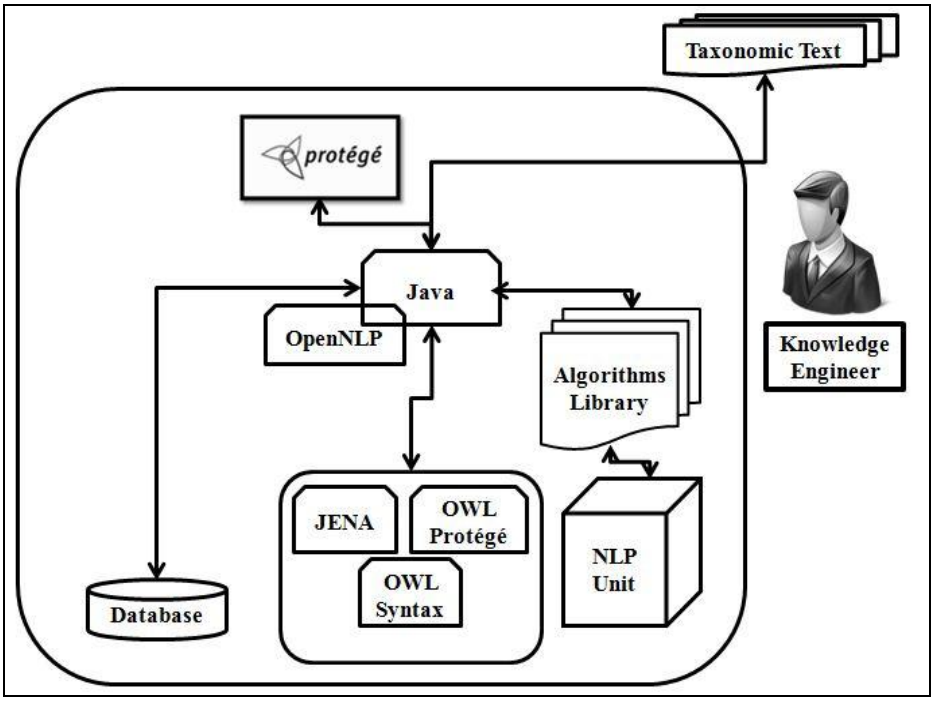


Fig.2 Representation of the process flow of software development using developed framework

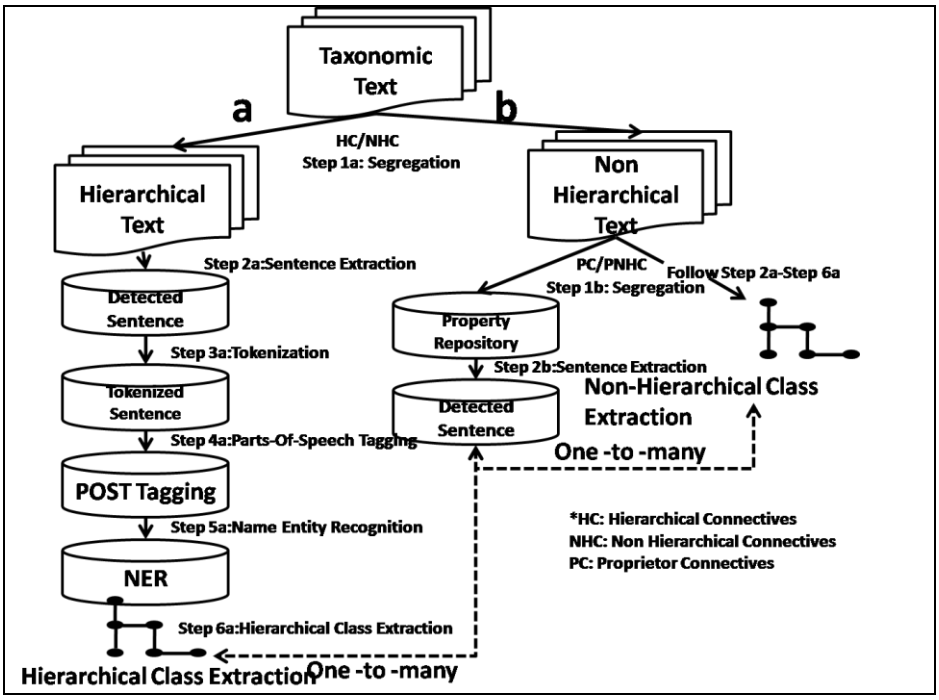


Fig.3 Entity Relationship Diagram of Developed System

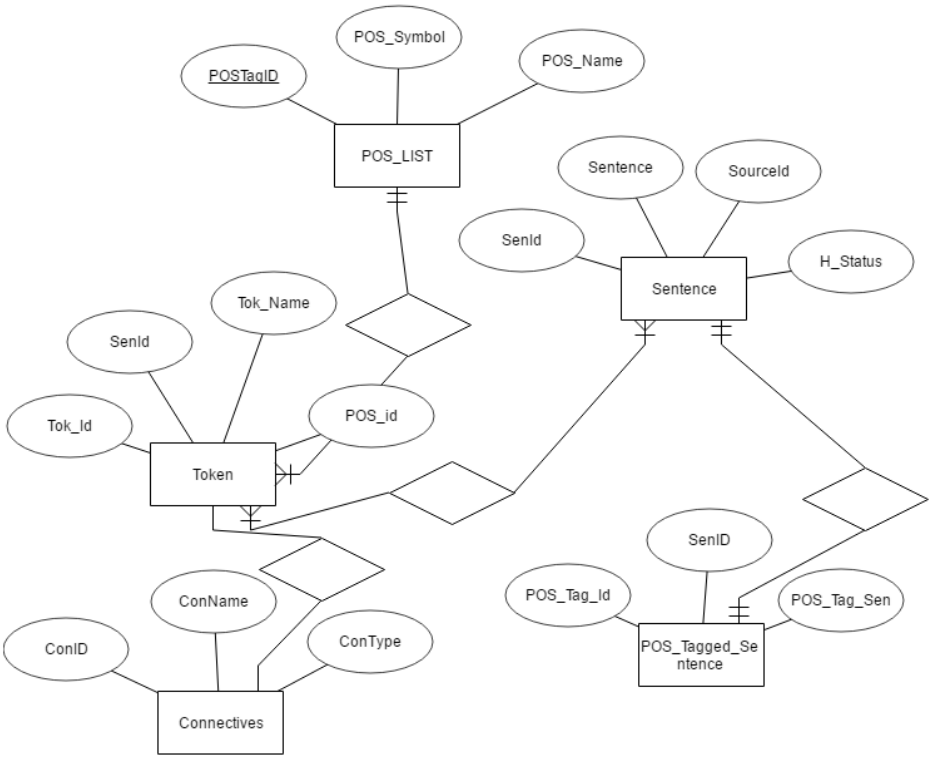


Fig.4 Schematic Class Diagram of the Developed System

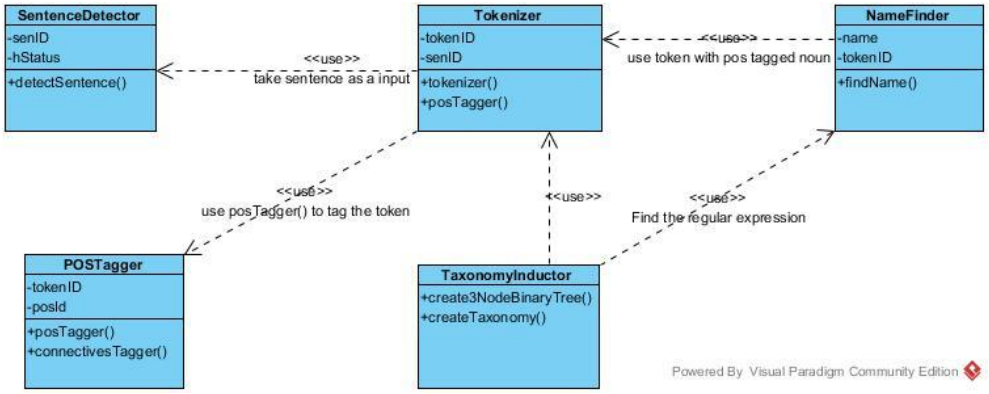


Fig.5 Classification of sentences into Hierarchical and Non-hierarchical sentences

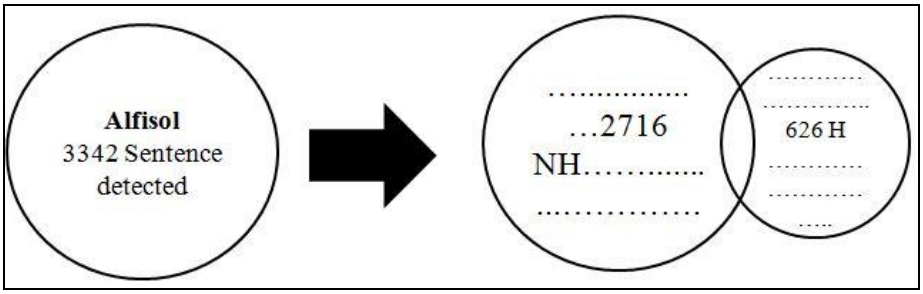


Fig.6 Representation of the three node parent child relationship of Aqualf and Alfisol from sentence 1 and 2

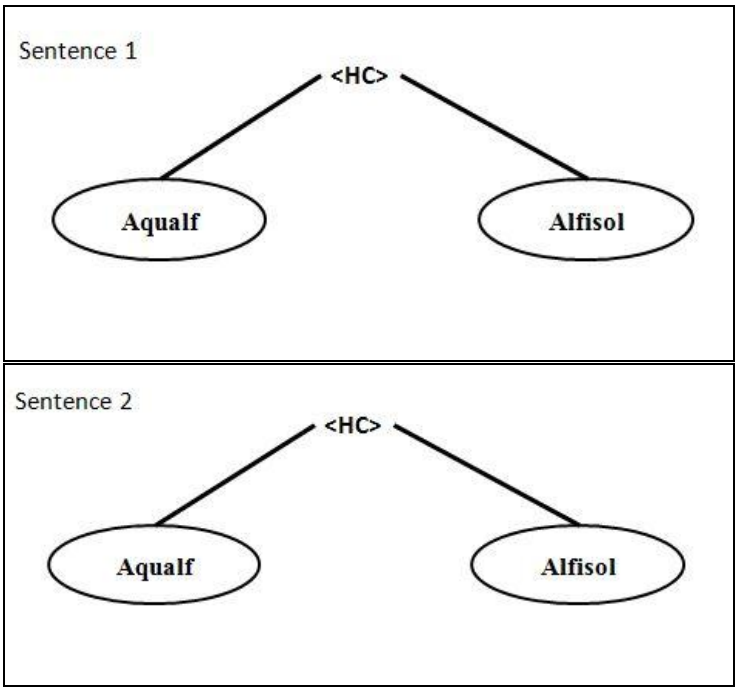


Fig.7 Equality relationship between Aquartic Cromic Haploudalf and Typic Haploudalf from sentence 134

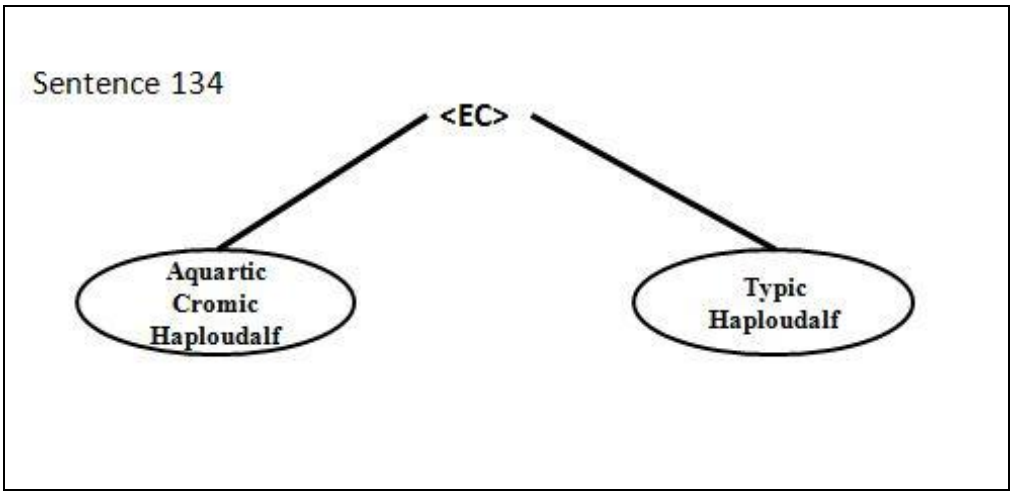


Fig.8 Identification of the classes with the parent child relationship of the Taxonomic class up to sub group

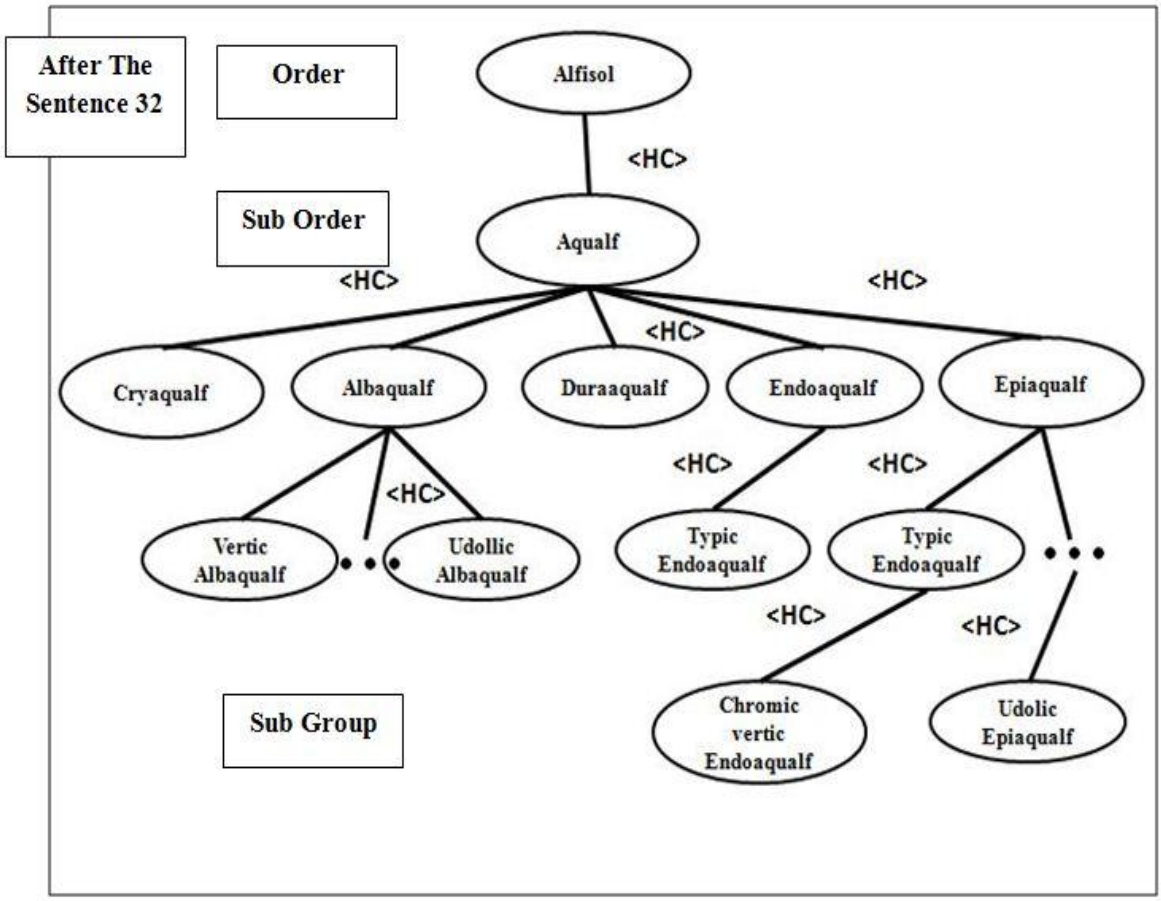
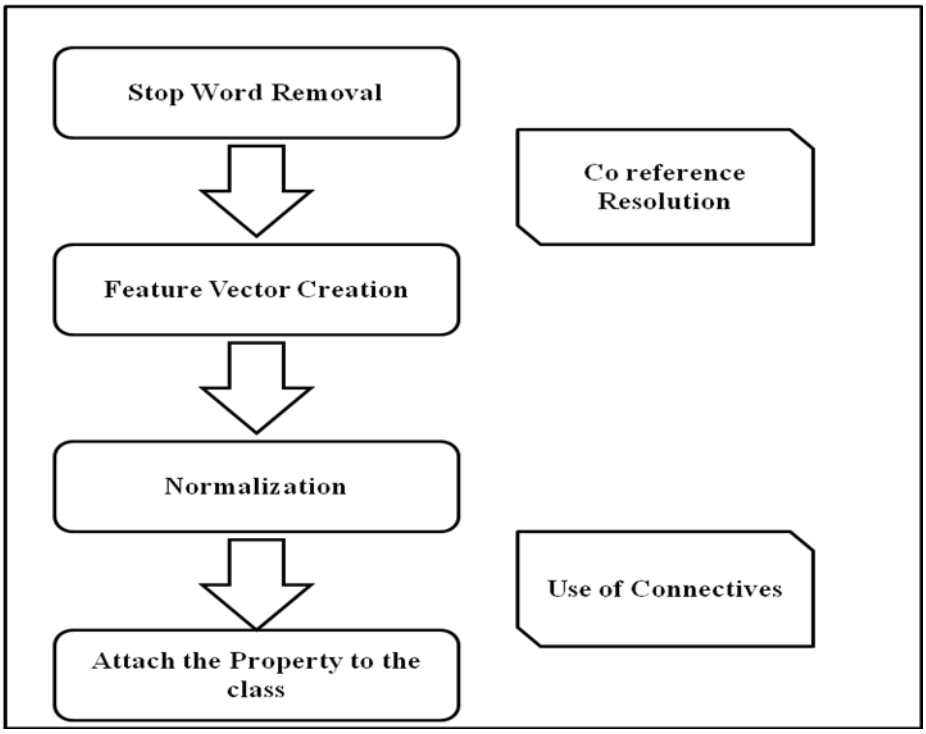


Fig.9 Flow diagram of the connective based property attachment to the Extracted class of soil taxonomy



The figure 7 shows the <EC> i.e. the equality connectives the root node of another three node binary tree developed from the sentence number 134. In this situation the binary tree depicts a little different from the previous two. Although the root node is the Equality node so the left child and right child will belong to the same super class i.e. these two node Aquartic Cromic Hapludalf and the typic hapludalf belongs to the same parent that would already be identified from the previous traversal of the sentence.

From the above discussion it is very much clear that in the three nodes binary three is the root node or the connectives has very important role to develop the whole taxonomy tree from the taxonomic text.

After creation of the entire binary tree we have plugged binary tree superimposing sub routine to the developed binary tree. Figure 8 shows that after traversing 32 three node binary tree the algorithm already identifies the level available in the input text. More elaborately we can say the USDA soil taxonomy, 2010 contains the taxonomic hierarchy up to the sub group level. Our algorithm all ready identify the levels after traversal of some of the binary tree.

Property Extraction

Previously, we have described the schematic process flow of the software development. The text used in experimentation has been segregated into two classes hierarchical and non-hierarchical. The hierarchical text used for extract the parent child relationship in the taxonomic text. The non-hierarchical text is used for the property extraction.

Figure 9 describe the step by step process for property identification of a particular class that is already been identified in the previous section.

In this work flow, first we have removed all the stop word from the text and then created the feature vectors and co-reference resolution. After that we normalize the vector with the identified classes. Identified property is attached to the particular class. We have identified the properties by the property base connectives.

In this research work we have inducted the taxonomy from USDA soil taxonomy, as well as extracted the property. More importantly we have done these tasks which are based on regular expression and with the help of connectives**.** Connective based taxonomy and property extraction helps us not to use huge corpus behind the text processing. In future the framework can be used for other domains in taxonomic text. It may also be extended for the other aspects of ontology learning like identification of axioms and constraints.

Acknowledgement

First author of the article gratefully acknowledge the INSPIRE Fellowship provided by Department of Science and Technology, New Delhi

References

Bedi, P., and Marwaha, S. 2004. Designing Ontologies from Traditional Taxonomies. In the Proceedings of International Conference on Cognitive Science, Allahabad. India.

Dietz, E. A., Vandic, D., Frasincar, F. 2012. Taxolearn: A semantic approach to domain taxonomy learning. In *Web Intelligence and Intelligent Agent Technology (WI-IAT).* 1: 58-65

Fountain, T., Lapata, M. 2012. Taxonomy induction using hierarchical random graphs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 466-476.

Galitsky, B., A. 2013. Transfer learning of syntactic structures for building taxonomies for search engines. *Engineering Applications of Artificial Intelligence*. 26(10): 2504-2515.

Hosny, M., M., El-Beltagy, S., R., and Allam, M., E. 2015. Unsupervised Data Driven Taxonomy Learning. In *2015 First International Conference on Arabic Computational Linguistics.* 9-14.

Kang, Y. B., Haghigh, P., D., and Burstein, F. 2014. C Finder: An intelligent key concept finder from text for ontology development. *Expert Systems with Applications*. 41(9): 4494-4504.

Kang, Y., B., Haghigh, P., D, and Burstein, F. 2016. TaxoFinder: A Graph-Based Approach for Taxonomy Learning. IEEE Transactions on Knowledge and Data Engineering. 28(2): 524-536

Knijff, J., De, Frasincar, F., and Hogenboom, F. 2013. Domain taxonomy learning from text: The subsumption method versus hierarchical clustering. *Data & Knowledge Engineering.* 83: 54-69.

Liu, K., Mitchell, K., J., and Chapman, W., W., Savova, G, K., Sioutos, N., Rubin, D., L., and Crowley, R., S. 2013. Formative evaluation of ontology learning methods for entity discovery by using existing ontologies as reference standards. 52(4): 308-16.

Liu, X., Song, Y., Liu, S., and Wang, H. 2012. Automatic taxonomy construction from keywords. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1433-1441.

Medelyan, O., Manion, S., Broekstra, J., Divoli, A., Huang, A., L., and Witten, I., H. 2013. Constructing a focused taxonomy from a document collection.

In Extended Semantic Web Conference. 367-381.

Meijer, K., Frasincar, F., and Hogenboom, F. 2014. A semantic approach for extracting domain taxonomies from text. *Decision Support Systems*. 62:78-93.

Ponzetto, S., P., and Strube, M. 2011. Taxonomy induction based on a collaboratively built knowledge repository *Artificial Intelligence*. 175(9-10):1737-1756

Rios-Alvarado, A., B., Lopez-Arevalo, I., and Sosa-Sosa, V., J. 2013. Learning concept hierarchies from textual resources for ontologies construction. *Expert Systems with Applications.* 40(15): 5907-5915.

Thukral, A., Jain, A., Aggarwal, M., and Sharma, M. 2018. Semi-automatic Ontology Builder Based on Relation Extraction from Textual Data. In: Bhattacharyya S., Chaki N., Konar D., Chakraborty U., Singh C. (eds) Advanced Computational and Communication Paradigms. Advances in Intelligent Systems and Computing. Springer, Singapore. vol.69.

Velardi, P., Faralli, S., and Navigli, R. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. 39(3): 665-707

Yang, H. 2012. Constructing task-specific taxonomies for document collection browsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 1278-1289

How to cite this article:

Chandan Kumar Deb, Madhurima Das and Sudeep Marwaha. 2018. Building Soil Taxonomy Ontology by the Way of Connective Based Ontology Learning. *Int.J.Curr.Microbiol.App.Sci.* 7(09): 975-985. doi: https://doi.org/10.20546/ijcmas.2018.709.117