ICIC Express Letters
Part B: Applications
Volume **10**, Number **8**, August **2019**

ICIC International ©2019 ISSN 2185-2766
pp. **673–680**

SEMANTIC RELATIONS EXTRACTION FOR ONTOLOGY CONSTRUCTION: A SURVEY

Mohamed Ali[1], Said Fathalla[1,2], Mohamed Kholief[3] and Yasser Hassan[1]

[1]Faculty of Science
Alexandria University
22 El-Guish Road, El-Shatby, Alexandria 21526, Egypt
M_hassan@nbe.com.eg; y.fouad@alexu.edu.eg

[2]Smart Data Analytics (SDA)
University of Bonn
Regina-Pacis-Weg 3, Bonn 53113, Germany
Sm_fathalla@alex-sci.edu.eg

[3]College of Computing and Information Technology
Arab Academy for Science, Technology and Maritime Transport
P. O. Box 1029, Abo Qir, Alexandria 21937, Egypt
kholief@aast.edu

Received December 2018; accepted March 2019

Abstract. *The objective of learning non-taxonomic relations of ontologies is the automatic extraction of all possible semantic relationships between ontology concepts in a particular domain. The discovering and labeling of non-taxonomic relations are the most difficult and not well researched in the ontology learning process. This work presents the results of the review of the most recent approaches during the last decade, with a focus on the solutions they provide, employed techniques, positive and negative aspects, and the evaluation metrics used. Our goal is to provide researchers in this area a comprehensive understanding of the drawbacks of the current existing work, thereby encouraging further improvement of the research work in this area. Therefore, a set of recommendations for future research is proposed.*
Keywords: Taxonomic relations, Non-taxonomic relations, Ontology evaluation

1. **Introduction.** Learning non-taxonomic relationships of ontologies (LNTRO) is a subtask of ontology learning (OL), which can be defined as the process of the automatic or semi-automatic construction of ontologies from a given corpus of a specific domain [4]. Most studies in LNTRO focus on discovering non-taxonomic relations (NTR). Labeling of these relations is considered the most challenging task [1]. *Problem Statement.* The problem of LNTRO is how to automatically, or semi-automatically, extract semantic relationships from text, i.e., NTR. In this paper, we present a study of the most recent existing approaches and evaluation methods that are used for LNTRO from an unstructured data source. We would like to point out that there is a short review paper published in 2012, by Serra et al. [2]. Surprisingly, we observed that there are no comprehensive review papers about LNTRO since 2012, except our previous short review [3] in 2017. Differently, this paper mainly focuses on the analysis of the most recent approaches. Therefore, we undertook this study in which we present the results of a systematic review of *five* approaches that represent the most recent state-of-the-art for LNTRO. We qualitatively analyze these approaches (Table 1). The solutions they provide are discussed along with their respective positive and negative aspects. Our goal is to provide researchers with a comprehensive understanding of the recently existing work in LNTRO, thereby

DOI: 10.24507/icicelb.10.08.673

encouraging further experimentation and new approaches. *Research Question.* We aim at discussing and analyzing current existing methodologies and evaluation methods for LNTRO. To achieve this goal, we aim to answer the following general research question: *"How can non-taxonomic relations be learned from unstructured data sources?"* We divide this general research question into further three sub-questions: 1) what are the main tasks and the techniques used in each task of LNTRO? 2) what are the advantages and disadvantages of each approach for LNTRO? and 3) what evaluation metrics are used to evaluate them? In this study, the selected papers were analyzed, compared and unified with respect to the techniques used, advantages, disadvantages, and evaluation metrics. The rest of this paper is organized as follows. In Section 2, we present an overview of the LNTRO subtasks. In Section 3, we present the selected LNTRO state-of-the-art approaches. In Section 4, the evaluation techniques used for LNTRO are presented. In Section 5, we discuss the results. Finally, in Section 6, we conclude the major drawbacks in the literature on LNTRO.

2. **LNTRO Subtasks.** Non-taxonomic relations (NTR) are those relations that exist between any concept pairs in the ontology except the taxonomic relation (is-a relation). In fact, this problem of the automatic extraction of NTRs appears to be a more intricate task as it is less well-known what type and how many of those relations should be modeled in a particular ontology. The major concern in NTRs learning is relation extraction and labeling [4]. Current research on extracting NTRs is based on statistical and semantic analysis approaches. The process of LNTRO can be accomplished through the following five tasks: 1) *corpus construction*: selecting related documents about the domain of interest, it is usually a challenging task because the outcome of any LNTRO technique depends on its quality, 2) *corpus annotation*: natural language processing (NLP) techniques are used to annotate the corpus with additional information that is needed for the posterior steps, 3) *extraction of relations*: candidate relations could be identified from the annotated corpus by using information extraction techniques, 4) *refinement of the relations*: relations from the previous task should not be recommended to the specialist since there is usually a substantial amount of them that do not correspond to good suggestions. For this reason, machine learning techniques can be used, and 5) *relation ingestion*: data mining techniques are used to suggest, to the specialist, the best possible level in the input ontology hierarchy where to add the relations. Table 2 presents tasks accomplished by each of the selected *five* approaches, while Table 3 presents the phases of LNTRO and their employed activities in each of the state-of-the-art approaches. Due to space limitation, we will use the following abbreviations in the rest of this paper: C for a concept, R for a relation, S for a sentence, V for a verb, ARs for association rules, and SVO for Subject-Verb-Object.

3. **LNTRO State-of-the-Art Approaches.** The discovery of NTR is considered as the most intricate task as, in general, it is not known how many and what type of conceptual relationships should be modeled in a particular ontology. Various approaches are presented in the literature for the learning of semantic relations among ontology concepts. In this section, the current state-of-the-art approaches for LNTRO will be presented. Most of LNTRO techniques use ontology taxonomy as input and a domain corpus. These techniques suggest the best level in the hierarchy where to insert the semantic relations. Those that receive only the ontology concepts have the search space for relations reduced and have the potential of obtaining better results when compared to those that do not receive this input. Techniques that do not receive any of these sets as input often consider noun phrases as concepts. Techniques on LNTRO are usually evaluated comparing their results against reference ontologies or gold standard ontologies [5]. However, comparing

them when executed under similar conditions is work that still must be done. In the following, we review the state-of-the-art approaches for LNTRO as a subtask of the ontology learning process.

3.1. **LNTRO based on correlation search.** Wong et al. [4] proposed a multi-phase correlation search framework for NTR extraction. The proposed framework addresses two sub-problems that are identified in the literature of this field: relation extraction and relation labeling. The proposed framework accomplishes its tasks in several steps: 1) extract correlated concept pairs, 2) filter out correlated concepts that have a taxonomic relation between them using existing domain ontology and thus, the remaining set of concept pairs could serve as candidates for NTRs learning, and 3) a pattern-based linguistic approach is used for labeling candidate relations. A set of possible labels is given to each of the candidate relations and let domain experts choose the most appropriate one depending on the domain in which the ontology will be used. *Positive Aspects*. 1) It searches for correlated domain concept pairs where the individual concepts may be across multiple adjacent sentences instead of just a single sentence, 2) the use of association rule mining (ARM) allows the search for correlated concept pairs beyond a single-sentence window as opposed to linguistic approaches, 3) the system can identify valid NTRs without human involvement, and 4) derive new correlations between pairs of concepts from n-itemsets association rules where $n > 2$. *Negative Aspects*. 1) The approach needs human intervention in the process of identifying the relevant labels for the candidate NTR, 2) it requires related domain ontology as a part of the learning process to identify those correlated concepts that have a taxonomic relationship between them, and 3) the results depend on the quality of the input related domain ontology and to which degree it covers the domain concepts of discourse.

Table 1. Selected approaches: Input, output, and language

Approach	Domain	Lang.	Input	Output
[4]	Marine Biology	English	Text documents in PDF format.	List of candidates labeled-NTRs.
[6]	Domain-independent	English	Domain ontology, domain-specific corpus of texts.	Enriched domain ontology by NTRs.
[7]	Domain-independent	English	Web pages belong to the domain.	An ontology with taxonomic and NTRs.
[8]	Wikipedia-DBpedia dataset	English	Seed ontology and homogeneous collections of text.	An ontology with extended instances, classes, taxonomic and NTRs.
[9]	Financial	English	Web corpus files.	An ontology with taxonomic and NTRs.

Table 2. Tasks accomplished by each approach

Approach	Extract concepts	Learn TR	Learn NTR		Evaluated
			Discovering	Labeling	
[4]	√	×	√	√	√
[6]	√	×	√	√	√
[7]	√	√	√	√	√
[8]	√	√	√	×	√
[9]	√	√	√	√	√

3.2. **LNTRO based on the unsupervised approach.** Ribeiro [6] proposed a framework for enriching ontologies by extracting NTRs given a domain ontology and a domain specific corpus. One key feature of the proposed framework is that the domain ontology which is used in the relation extraction is also the target ontology to be enriched. The conceptual architecture of the proposed framework comprises four components: preprocessing, predicate identification, relation occurrences identification, and association mining. A case of study for Tennis sports domain has been introduced. In the preprocessing, unstructured documents are converted to a suitable format for further processing. In predicate identification, verbs are identified from text documents using the POS tagging step in the pre-processing component. At this stage, the identified verbs, present in the sentences selected in the previous component are counted as a predicate in set P, where P is a set of predicates that contains all relevant verbs, then calculates the weight for each group of predicates present in P'. In association mining, once a set of candidate relations between concepts is available, they should be validated before suggesting them to enrich the ontology. To achieve this, above expectation measure is used [10]. *Positive Aspects*. 1) It is a fully unsupervised approach, 2) it supports the pronoun resolution, 3) it allows for the combination of domain-independent linguistic techniques for the extraction of relation candidates (based on the highest term frequency-inverse document frequency (TF-IDF) value), and 4) it is domain-independent. *Negative Aspects*. 1) Only information explicitly present in the corpus is extracted, no inference for implicit data, 2) the extraction of NTRs is based on the concepts presented in the domain ontology, that means concepts outside the ontology will be ignored, 3) the proposed framework is based on the linguistic structure of the English language, i.e., language-dependent approach, 4) one of the limitations of this framework is that it does not support n-ary relations, and 5) the required related domain ontology may not be readily available.

Venu et al. [7] proposed an unsupervised approach for building domain ontologies without the use of any annotated resource. After corpus collection and term extraction, the NTR extraction process begins. In NTR extraction, two techniques have been used: triplet extraction in which Rusu's triple algorithm [11] is used to extract triples and association rule mining in which Apriori algorithm [12] is used for finding the NTRs between terms. Association rules which satisfy a threshold (confidence score) are selected to be learned. *Positive Aspects*. 1) Ontology was automatically built from scratch without supervision, 2) authors developed an iterative focused crawler for the collection of domain corpora for ontology construction, and 3) building ontology from scratch including contacting required corpus and NTR. *Negative Aspects*. 1) It needs more implementation and evaluation techniques to clarify its efficiency, and 2) it uses a combination of existing approaches in the learning process of ontology components.

3.3. **LNTRO based supervised approach.** Starc and Mladenić [8] proposed a novel approach to joint learning of ontology and semantic parser from text. The method is based on semi-automatic induction of a context-free grammar from the semantically annotated text. The grammar parses the text into semantic trees. Both, the grammar and the semantic trees are used to learn ontology on several levels – classes, instances, taxonomic and NTRs. The relations have been learned from semantic trees. Given a dataset of positive relation examples that represent one relation type, e.g., birthplace, the goal is to discover new unseen relations. This method is based on the assumption that a relation between entities is expressed in the shortest path between them in the semantic tree [13]. The input for the training process is the sentences in layered representation, corresponding parse trees, and relation examples. Given a relation from the training set, we first try to identify the sentence containing each entity of the relation. The relation can have one, two, or even more entities. Each entity is matched to the layer that corresponds to the entity type. For example, strings are matched to the lexical layer; ontology entities are

matched to the layer containing such entities. *Positive Aspects*. 1) It can be applied to many languages as it discovers the grammar and builds a parse tree from the input corpus, and 2) it builds ontology components from scratch. *Negative Aspects*. 1) It is designed for homogeneous collections of text to avoid data redundancy, this restriction considered as a limitation of this approach, 2) it is text-driven, semi-automatic and based on grammar induction, and 3) it needs human intervention.

Table 3. Phases of LNTRO and their employed activities for each approach

Ref	Corpus preprocessing		NTR extraction		Relation labeling/ Ingestion
	Construction	Preparing	Discovering	Refinement	
[4]	Fisheries Oceanography journal in the PDF format.	Convert PDF files to text files, remove unessential data, combine files into one text and split it to sentences.	Patterns that express NTRs are discovered, and GATE components are used in verb extraction.	The C-V-C triples ordered by numerical measure and top ones are selected as candidates for relation labels.	A pattern-based linguistic approach is used to provide a meaningful label to the NTRs.
[6]	Unstructured documents that represent domain-specific corpus.	Segment documents into sentences using punctuation marks like an exclamation mark, full stop, and question mark.	Identified verbs are counted as a predicate in set P, predicates with a similar meaning are grouped together, and group weights are calculated by TF-IDF.	If a subject or object were not labeled as ontological concepts, they are ignored during triplet construction.	The SVO method chooses V between the two C present in a concept pair for which its group has the highest TF-IDF value.
[7]	Seed URLs are given to the iterative focused crawler to produce Web pages relevant to the domain.	Web crawler downloads the contents which are pertinent to the domain. URL satisfying the relevance score is added to the URL queue.	ARM on triples is used to extract NTR between two correlated concept pair and Apriori algorithm is used for frequent itemset generation.	Association rules are filtered from frequent itemsets and association rules which satisfy a suitable confidence score are selected.	The extracted relations consist of property, domain, and range. The property represents the label of the NTR.
[8]	Homogeneous collections of text.	Annotate texts including annotations with the concepts from the existing ontology.	Give sentences, corresponding parse trees, and relation examples for the training process.	Not approached	Not approached
[9]	Collect a large number of web corpus files.	K-means partition corpus into k clusters. Indexing process is used to index the clustered files.	Extract all S where C is found. For each S, discover all R using one of Open IE algorithms.	Each tuple is judged as related based on whether the existence of C in one of the extracted arguments.	Open IE returns verbs that represent labels of the extracted NTR.

3.4. **LNTRO based Open Information Extraction.** Esserhrouchni et al. [9] proposed a new methodology for learning NTRs and building financial ontology from scratch. This technique is based on using, integrating and adapting Open Information Extraction algorithms to extract and label domain relations between concepts. The proposed process for learning non-taxonomic domain relationships with Open IE tools is performed in three steps: 1) for each concept of the taxonomy, all the sentences are extracted from the corpus where a concept is found, 2) for each extracted sentence, all possible relations are discovered using one of the Open IE algorithms, the output is a set of relational tuples $< Arg1, Rel, Arg2 >$ that describe the sentence verb relation (Rel) and its arguments (*Arg1* and *Arg2*), and 3) each resulted tuple is judged as related to the studied domain or not, based on whether it contains the concept C in one of the extracted arguments. The selected relations are incorporated into the resulting ontology. This process is repeated until all concepts of the taxonomy are processed. *Positive Aspects*. 1) Open IE algorithms are used to extract relations from a large text corpus, 2) Open IE systems facilitate domain independent discovery of relations as they extract all possible relations

without any prerequisite or restriction, 3) the indexing process as a corpus pre-processing step allows an efficient and fast retrieval of the needed information in the next stages in the learning process, and 4) this work available in the online publication as a web application. *Negative Aspects*. 1) The need to increase the size of the finance corpus in order to build a richer ontology for finance domain, and 2) applied only in financial domain. Table 4 summarizes the techniques used by each of the selected approaches and the task in which these techniques are used. For instance, in most of these approaches, association rule mining algorithms are used to identify/suggest NTR from a text corpus, while all of them use NLP techniques, such as POS tagging, sentence splitter, and tokenization to prepare the input text processing it.

Table 4. Employed techniques and their tasks in each approach

Ref	Year	Employed techniques and their tasks		Learning
[4]	2014	NLP	Remove unnecessary information and split the text into a list of sentences.	Semi-automatic
		DM	*Mining association rules approach* is used to extract correlated concepts, then filter them using association rule mining.	
[6]	2014	NLP	Sentence splitter, POS-tag, and named entity recognition.	Semi-automatic
		ML	*An unsupervised approach* is used to extract NTRs based on the concepts present in the domain ontology in one pass, rather than extracting them in different iterations.	
[9]	2015	NLP	Tokenization, normalization, lemmatization and stopword removal.	Automatic
		IE	The *Open Information Extraction approach* discovers NTRs and performs an iterative mining algorithm that constructs the ontology.	
[7]	2016	NLP	Remove stop words, and then tokenize documents into sentences.	Automatic
		DM	*An unsupervised approach*, such as association rule mining, is used to extract NTR.	
[8]	2017	NLP	Text annotation, shallow NLP tools, like sentence splitting, word tokenization, named entity recognition.	Semi-automatic
		ST	*Supervised Approach* is used to learn NTRs from semantic trees, given a dataset of positive relation examples that represent one relation type.	

4. **Evaluation Techniques for LNTRO.** Evaluation of ontology learning systems, in general, is the assessment of the resulting ontologies which lead to guiding and refining the learning process. It is still an open problem and there is still little research in this direction. According to [5], the resulting ontology from the learning process could be evaluated by 1) using it in an executable application; 2) by domain experts or even by 3) comparing it with a predefined reference ontology, such as in [14] (i.e., gold-standard based evaluation). The focus of OL evaluation is to determine whether the terms/relations used in the learning process are correct. Lexical precision (LP) is the fraction of retrieved concepts and relations that are relevant, while lexical recall (LR) is the fraction of relevant items retrieved by the system. F-measure is a mix of both LP and LR by calculating the harmonic mean of them. Hlomani and Stacey [15] proposed a four-layered metric suite for ontology evaluation which is inspired from the metric suite for ontology auditing, which

is proposed in [16]. The metric suite compromises a set of evaluation criteria, such as accuracy, completeness, conciseness, consistency, computational efficiency, adaptability, and clarity. The major aim is to evaluate an ontology according to the decision of experts on whether to reuse the ontology or not. As shown in Table 5, two approaches use gold standard evaluation to compare the resultant enriched ontology to a gold standard one by a domain expert and two approaches use LP and LR. One approach uses metric-based and analysis of variable tree each. In addition, some approaches, such as the one proposed in [9], use both gold standard evaluation and precision and recall measures. On the other hand, metric-based, such as inheritance and class richness evaluation and analysis of variable tree evaluations are rarely used.

Table 5. Evaluation metrics used by each approach

Approach	**Evaluation metrics**			
	Gold standard	LP and LR	Metric-based	Analysis of variable tree
[4]	√			
[6]		√		
[7]			√	
[8]				√
[9]	√	√		

5. **Discussion.** Here, we summarize the main drawbacks of most of the state-of-the-art approaches for LNTRO: 1) they are based on the analysis of syntactic structures of specific language and dependencies among concepts existing in a domain-specific text corpus which means that all those approaches are domain and language dependent, 2) they are semi-automatic, i.e., need human intervention to suggest the possible relationships among concepts, validate the extracted relations, and select the most appropriate labels for the extracted relations, 3) they lack rigid measures for evaluation, 4) they do not appropriately solve the labeling problem, and 5) they neglect the importance of the appropriate level of abstraction. *Recommendations*. Based on the results of this review, we propose the following recommendations for future research in this area: 1) analyze the impact of ontology design that is given as input in the extraction process of NTR, ontologies which are more carefully designed should thus yield better extraction results, 2) work is needed to enhance and optimize the process of constructing and preparing input corpus for the learning process, as the quality of the outcome of most LNTRO approaches highly depends on the quality of the input corpus, 3) approaches that can deal with the noise that might arise when the input data crowded from web documents are needed, which can be applied to a wide possible range of situations, and 4) set of techniques and measures are needed to enhance, optimize, evaluate and determine the efficiency of the input corpus and the resultant ontology.

6. **Conclusion.** This paper presents, to the best of our knowledge, the most comprehensive systematic review of the state-of-the-art approaches for LNTRO and its subtasks. *Five* approaches have been described along with their positive and negative aspects. The goal of this survey is to obtain a clear understanding of the problem of LNTRO and its sub-problems which have been addressed in the literature. As our literature review reveals, this research area still needs a lot of work to produce fully automatic approaches that are capable of LNTRO in an efficient and independent manner. After reviewing the literature, we have identified the main drawbacks of most of the state-of-the-art approaches for LNTRO and consequently, based on these drawbacks, we have proposed a set of recommendations for the future research in this area of research. The most outstanding

conclusion is that most of the state-of-the-art approaches: 1) focus on learning one ontology component and ignore the other components, 2) learn from a given corpus in a specific language with a relevant linguistic characteristic that differs from the other languages, so it cannot be used for the learning process in another language, i.e., language-dependent, 3) need human intervention, and 4) domain-dependent; i.e., they are tailored for a specific domain, but it might fail with other domains. However, the key problem with much of the literature on LNTRO is that the problem of semantic ambiguity presented in natural language resources is not taken into consideration. Furthermore, one of the major drawbacks in the literature on LNTRO is that they are domain-dependent and language-dependent. Therefore, there is a pressing need to develop language-independent approaches or at least able to use learned NTRs in English to learn the corresponding to it in another language.

REFERENCES

[1] A. Maedche and S. Staab, Discovering conceptual relations from text, *Proc. of the 14th Eur. Conf. Artif. Intell.*, pp.1-17, 2000.

[2] I. Serra, R. Girardi and P. Novais, The problem of learning non-taxonomic relationships of ontologies from text, *Adv. Intell. Soft Comput.*, vol.151, pp.485-492, 2012.

[3] M. Ali, S. Fathalla, M. Kholief and Y. F. Hassan, The problem learning non-taxonomic relationships of ontologies from unstructured data sources, *2017 23rd IEEE International Conference on Automation and Computing: Addressing Global Challenges through Automation and Computing*, 2017.

[4] M. K. Wong, S. S. R. Abidi and I. D. Jonsen, A multi-phase correlation search framework for mining non-taxonomic relations from unstructured text, *Knowl. Inf. Syst.*, vol.38, no.3, pp.641-667, 2014.

[5] K. Dellschaft and S. Staab, On how to perform a gold standard based evaluation of ontology learning, *International Semantic Web Conference*, pp.228-241, 2006.

[6] N. L. Ribeiro, *Extraction of Non-Taxonomic Relations from Texts to Enrich a Basic Ontology*, Master Thesis, Instituto Superior Tcnico, 2014.

[7] S. H. Venu, V. Mohan, K. Urkalan and T. V. Geetha, Unsupervised domain ontology learning from text, *International Conference on Mining Intelligence and Knowledge Exploration*, pp.132-143, 2016.

[8] J. Starc and D. Mladenić, Joint learning of ontology and semantic parser from text, *Intell. Data Anal.*, vol.21, no.1, pp.19-38, 2017.

[9] O. El Idrissi Esserhrouchni, B. Frikh and B. Ouhbi, Learning non-taxonomic relationships of financial ontology, *Proc. of the 7th Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag.*, Lisbon, Portugal, pp.479-489, 2015.

[10] M. Kavalec and A. Maedche, Discovery of lexical entries for non-taxonomic relations in ontology learning, *The 30th Conf. Curr. Trends Theory Pract. Comput. Sci.*, Měřín, Czech Republic, 2004.

[11] D. Rusu, L. Dali, B. Fortuna, M. Grobelnik and D. Mladenić, Triplet extraction from sentences, *Proc. of the 10th Int. Multi-Conference Inf. Soc.*, 2007.

[12] R. Agrawal and R. Srikant, Fast algorithms for mining association rules, *The 20th Int. Conf. Very Large Data Bases*, 1994.

[13] R. A. Frost and R. Hafiz, A new top-down parsing algorithm to accommodate ambiguity and left recursion in polynomial time, *ACM SIGPLAN Not.*, 2006.

[14] M. Ali, S. Fathalla, S. Ibrahim, M. Kholief and Y. Hassan, Cross-lingual ontology enrichment using multi-agent architecture, *Procedia Computer Science*, vol.137, pp.127-138, 2018.

[15] H. Hlomani and D. Stacey, Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey, *Semant. Web J.*, vol.1, no.5, pp.1-11, 2014.

[16] A. Burton-Jones, V. C. Storey, V. Sugumaran and P. Ahluwalia, A semiotic metrics suite for assessing the quality of ontologies, *Data and Knowledge Engineering*, vol.55, no.1, pp.84-102, 2005.