CONCEPT AND RELATION EXTRACTION FRAMEWORK FOR ONTOLOGY LEARNING

FATIMA NADEEM SALEM AL-ASWADI

UNIVERSITI SAINS MALAYSIA

2023

CONCEPT AND RELATION EXTRACTION FRAMEWORK FOR ONTOLOGY LEARNING

by

FATIMA NADEEM SALEM AL-ASWADI

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

August 2023

DEDICATION

I dedicate this research to my parents' souls, who have constantly instilled hope in my soul through their irreplaceable giving. I am forever indebted to them for having made me who I am.

In addition, I dedicate this research to every ambitious person who is interested in upscale science and genuinely seeks to reach the goal for the benefit of humankind.

ACKNOWLEDGEMENT

First and foremost, praise be to Allah, Lord of the Worlds, Most Gracious, Most Merciful, for His gracious care and support, which have enabled me to complete this thesis. I would have never been able to get through my PhD journey if not for His mercy. I am grateful to Allah for His blessings and guidance at each step in my life.

I feel honoured to express my sincere appreciation and gratitude to my supervisor Associate Professor Dr. Chan Huah Yong, and co-supervisor Associate Professor Dr. Gan Keng Hoon, for their wise guidance, encouraging feedback, and invaluable help during my PhD journey. They are my primary resources for getting my scientific questions answered. I hope that I could be as lively, enthusiastic, and energetic as they are. Also, I would like to express my heartfelt thanks and deepest gratitude to all my family members: my father and mother (may Allah have mercy upon them), my brothers, sisters, uncles, aunts, cousins, and friends, especially my best friend Wafa'a Alma'aitah, for their unconditional love, unlimited help and for their faith in me. I would not have been able to go through everything without their constant encouragement and support.

I would like to acknowledge that this research was partially supported by Universiti Sains Malaysia (USM) under Grant-Account no. 203.PKOMP.6777003. I would like to express my thanks to USM, especially the School of Computer Sciences, for their considerable efforts in guiding the students. My sincere thanks and appreciation also go to Hodeidah University, especially to the Deanship of Graduate Studies and to the Faculty of Computer Science & Engineering, for providing me with all the necessary facilities and for the scholarship. Finally, I express profound thanks to all who have contributed, directly or indirectly, to this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
ASIUM	Acquisition of Semantic knowledge Using Machine
BCV-BOW	Binary Concept-Verb Bag-of-Words
BERT	Bidirectional Encoder Representations from Transformers
BOW	Bag-of-Words
BP-NN	Back Propagation Neural Network
BP-NN	Back Propagation Neural Network
CBOW	Continuous Bag-of-Words
CE-stopwords	Concepts Extraction Stopwords
CNN	Convolutional Neural Network
CRCTOL	Concept-Relation-Concept Tuple based Ontology Learning
CREF	Concept and Relation Extraction Framework
CRF	Conditional Random Field
DBN	Deep Belief Network
DC	Domain Consensus
DL	Deep Learning
DLO	Deep Learning Ontology
DNN	Deep Neural Network
DODDLE	A Domain Ontology rapiD DeveLopment Environment
DR	Domain Relevance
DTC	Domain Time Concept
DTR	Domain Time Relevance
DTR-DCEM	DTR Developed Concept Extraction Method

EAEDB	Entity Attribute Extraction Based on Deep Belief Network
GNN	Graph Neural Network
HMM	Hidden Markov Model
IG	Information Gain
IR	Information Retrieval
KIF	Knowledge Interchange Format
KM	Knowledge Management
LSTM	Long short-term memory
ML	Machine Learning
MLP	Multilayer Perceptron
NB	Naïve Bayes
NER	Named Entity Recognition
NLP	Natural Language Processing
OL	Ontology Learning
OOV	Out-of-Vocabulary
OWL	Web Ontology Language
POS	Part-Of-Speech
QA	Question-Answering
RBM	Restricted Boltzmann machine
RDF	Resource Description Framework
Re-DBN	Relu Deep Belief Network
Re-DDBN	Relu Dropout Deep Belief Network
ReD-RBM	Relu Dropout Restricted Boltzmann machine
Relu	Rectified Linear Activation Function
RNN	Recurrent Neural Network

Sig-DDBN	Sigmoid Dropout Deep Belief Network
SVM	Support Vector Machine
SYNDIKATE	SYnthesis of DIstributed Knowledge Acquired from TExts
TF-IDF	Term Frequency Inverse Document Frequency
TF-IDF-BOW	Advanced Bag-of-Words
TVoted	Intersection Voting Scheme
UMLS	Unified Medical Language System
YAGO	Yet Another Great Ontology

LIST OF APPENDICES

RANGKA KERJA PENGEKSTRAKAN KONSEP DAN HUBUNGAN UNTUK PEMBELAJARAN ONTOLOGI

ABSTRAK

Pengekstrakan pengetahuan yang berharga dan mewakilinya dalam bentuk yang boleh difahami oleh mesin dianggap sebagai salah satu cabaran utama bidang web semantik dan kejuruteraan pengetahuan. Pertumbuhan pesat data berbentuk teks iringi oleh peningkatan permintaan ontologi. Pembelajaran Ontologi (OL) daripada teks ialah proses yang bertujuan untuk mengekstrak secara automatik atau separa automatik dan mewakili pengetahuan daripada teks ke dalam bentuk yang boleh difahami oleh mesin. Ontologi ialah skema teras yang mewakili pengetahuan sebagai set konsep-konsep dan hubungan mereka dalam suatu domain. Pengekstrakan konsep-konsep dan hubungan mereka adalah tulang belakang sistem OL. Sistem-sistem OL yang sedia mempunyai banyak batasan dan kelemahan, seperti tidak cekap untuk mengekstrak konsep yang berkaitan terutamanya, untuk set data bersaiz besar; bergantung pada sejumlah besar corak yang telah ditentukan untuk pengekstrakan perhubungan, dan hanya boleh mengekstrak perhubungan yang beberapa jenis. Dalam tesis ini, satu rangka kerja yang dikenali sebagai Rangka Kerja Pengekstrakan Konsep dan Hubungan (CREF) telah dicadangkan. Ia terdiri daripada empat peringkat utama: peningkatkan kaedah pra-pemprosesan, pembangunan konsep kaedah pengekstrakan, pencadangkan pendekatan perwakilan teks baharu untuk perhubungan, dan penambahbaikan kaedah pengekstrakan hubungan. Peringkat pertama melibatkan pencadangan kata henti Pengekstrakan Konsep (CE-stopword) baharu untuk penerbitan saintifik manakala peringkat kedua melibatkan pengenalan metrik Perkaitan Masa Domain (DTR) baharu dan pencadangan Kaedah Pengekstrakan

Konsep Dibangunkan berdasarkan DTR yang dipanggil (DTR-DCEM). Peringkat ketiga melibatkan pencadangan pendekatan perwakilan teks baharu untuk pengekstrakan dan pengelasan hubungan yang dikenali sebagai Beg-Perkataan Konsep-Kata Kerja Binari (BCV-BOW). Akhir sekali, peringkat ke-empat melibatkan dua langkah; langkah pertama ialah penyelarasan *deep belief network (DBN)* untuk pengekstrakan hubungan menggunakan strategi penciciran; model ini dkenali sebagai Re-DDBN. Manakala langkah kedua melibatkan pembangunan pengekstrakan dan pengelasan hubungan menggunakan Re-DDBN. Eksperimen kepelbagaian telah dijalankan untuk menilai keberkesanan setiap kaedah yang dicadangkan pada setiap peringkat serta keberkesanan rangka kerja yang dicadangkan secara keseluruhan. Keputusan eksperimen menunjukkan bahawa kaedah dan rangka kerja yang dicadangkan mempunyai prestasi yang lebih baik dalam pengekstrakan konsep dan perhubungan yang berkaitan untuk pembinaan ontologi apabila disbanding dengan model yang dibentang adalam literatur erkini. Kaedah yang dicadangkan telah mengatasi model perbandingan seperti Text2Onto, Dev-Text2Onto, tf-idf-DCEM, DC-DR-Text2Onto, dan DC-DR-DCEM dalam tugas pengekstrakan konsep yang berkenaan, dengan peningkatan prestasi F1 antara 3.29% hingga 51.07%. Untuk tugas pengekstrakan dan pengkelasan hubungan, kaedah yang dicadangkan menunjukkan peningkatan prestasi, antara 9.33% hingga 35.84%, berbanding dengan model perbandingan seperti Mesin Vektor Sokongan (MVS) dan Naive Bayes (NB). Sama juga, penilaian keseluruhan menunjukkan bahawa kaedah yang dicadangkan telah mengatasi model perbandingan berdasarkan SVM atau NB dengan margin 3.87% hingga 12.65%. Tambahan pula, signifikasi statistik untuk kaedah yang dicadangkan telah dibuktikan dengan menggunakan ujian-t berpasangan (paired t-test).

CONCEPT AND RELATION EXTRACTION FRAMEWORK FOR ONTOLOGY LEARNING

ABSTRACT

Extracting valuable knowledge and representing it in a machine-understandable form is considered one of the main challenges of semantic web and knowledge engineering fields. The explosive growth of textual data is coupled with the increasing demand for ontologies. Ontology Learning (OL) from text is a process that aims to automatically or semi-automatically extract and represent the knowledge from text into the machine-understandable form. Ontology is a core scheme representing knowledge as a set of concepts and their relationships within a domain. Extracting the concepts and their relations is the backbone for an OL system. The existing OL systems have many limitations and drawbacks, such as not efficient for extracting relevant concepts especially, for large-length dataset; depending on a large amount of predefined patterns to extract relations, and extracting very few types of relations. In this thesis, a framework called Concept and Relation Extraction Framework (CREF) is proposed. It consists of four main stages: enhancing pre-processing method, developing methodology for the concept extraction task, proposing a new text representation approach for relations, and improving relation extraction method. The first stage involves proposing a new Concept Extraction stopwords (CE-stopwords) for scientific publications while the second stage involves introducing a new Domain Time Relevance (DTR) metric and proposing a Developed Concept Extraction Method based on DTR called (DTR-DCEM). The third stage involves proposing a new text representation approach for relation extraction and classification called Binary Concept-Verb Bag-of-Words (BCV-BOW). Finally, stage four involves two steps; the first one is configuring and tuning the deep belief network (DBN) for relations extraction using dropout strategies; this model is called Re-DDBN. While the second step involves developing relations extraction and classification using Re-DDBN. Extensive experiments were conducted to evaluate the effectiveness of each proposed method in each stage and to evaluate overall the effectiveness of the proposed framework. Experimental results confirm that the proposed methods and framework have significantly outperformed current state-of-the-art models in extracting the relevant concepts and relations for constructing ontologies. The proposed methods outperformed comparative models such as Text2Onto, Dev-Text2Onto, tf-idf-DCEM, DC-DR-Text2Onto, and DC-DR-DCEM in the task of relevant concepts extraction, with an improvement in F1 performance ranging between 3.29% and 51.07%. For the task of relations extraction and classification, the proposed methods exhibited improved performance, ranging from 9.33% to 35.84%, compared to comparative models like the Support Vector Machine (SVM) and Naive Bayes (NB). Similarly, the overall evaluation showed that the proposed method outperformed comparative ones, which used SVM or NB, by a margin of 3.87% to 12.65%. Moreover, the statistical significance of the proposed methods has been proved using paired t-test.

CHAPTER 1

INTRODUCTION

This chapter presents a background of ontology construction, ontology learning (OL), concepts extraction, and relations extraction for OL. In addition, the research motivation, questions, objectives, contributions, problem statement, scope and limitations, as well as thesis outlines are presented in this chapter.

1.1 Background of Ontology Construction

Ontology is considered one of the main cornerstones of representing knowledge in a more meaningful way on the semantic web. It represents knowledge in a structured, consistent, and understandable format for both computers and humans, thereby facilitating efficient retrieval, usage, storage, and maintenance of data. Usage of ontologies has proven to be beneficial and efficient in different applications, such as Information Retrieval (IR), Question-Answering (QA), Semantic Searching, Decision-Support, and Automated Fraud Detection (da Silva et al., 2020; Ergeta, 2019; Franco et al., 2020; Zou, 2020). There are two types of Ontology: (i) formal ontologies, which involve taxonomies, concepts with detailed relations between them, and the rules; (ii) informal ontologies that are created by user communities such as the internet encyclopedias (Astrakhantsev & Turdakov, 2013). This study focuses on formal ontology, so any indication of ontology in this thesis refers to formal ontology.

1.1.1 Ontology Formal Definition

According to W3C, *"Ontologies define the terms used to describe and represent an area of Knowledge"*. The ontology is a data model that represents the knowledge in a set of relevant concepts and the relations among those concepts within a domain. (Mishra & Jain, 2015). Zouaq (2011) defined the components of an ontology by the following tuple:

$$O =< C, H, Rr, A >$$

Where O represents ontology, C represents a set of classes (concepts), H represents a set of hierarchical links between the concepts (taxonomic relations), Rr represents the set of conceptual links (non-taxonomic relations), and A represents the set of rules and axioms (rules and axioms are out the scope of this thesis).

1.1.2 Ontology Construction

Ontology construction can be defined as an iterative process of creating an ontology from scratch or reusing an existing ontology for enriching or populating. The ontology construction process from text is a procedure that involves analyzing the collected text for a specific domain, identifying the relevant concepts and their relationships, mapping and representing the ontology by representation language [e.g. OWL (Web Ontology Language), RDF (Resource Description Framework)]. The ontology construction process may be done by one of the following three ways:

i. Manual construction: experts perform manual construction of ontology

ii. Cooperative construction: most or all tasks of the ontology construction system are performed or supervised by experts.

iii. (Semi-) Automatic construction: the ontology construction process is performed automatically with limited intervention by users or experts. Automatic construction means that the level of human intervention is slightly less than semi-automatic construction but does not mean fully automatic construction.

Manual or cooperative construction of ontologies is a time-consuming, extremely laborious, and costly process (Ma & Molnár, 2020; Maimon & Browarnik, 2015). In recent years, many approaches and systems that try to automate the construction of ontologies have been developed. It is worth mentioning that full automatic construction for ontology by a system is still a significant challenge, and it is not likely to be possible (Maimon & Browarnik, 2015; Siebra & Wac, 2023; Wong, Liu, & Bennamoun, 2012).

1.1.3 Ontology Learning

OL is a process of automatically or semi-automatically creating new ontology, or enriching, or sometimes populating an existing ontology; without or with the minimum human intervention (Sathiya & Geetha, 2018). OL from text is a process of acquiring knowledge from the text in a specific domain by applying natural language processing (NLP), data mining, and machine learning (ML) techniques (Gillani Andleeb, 2015; Narayanasamy, Kathiravan Srinivasan, & Chang, 2021). OL process includes five sub-tasks: terms extraction, synonyms discovery, concepts extraction, relations extraction, and rules or axioms extraction. These thesis contributions are conducted to develop concepts extraction and relations extraction methods for OL as they are the main parts of constructing ontologies.

There are many studies that compare between OL methods and techniques. It is worth remarking that the comparison between different OL methods is difficult because there is no much consensus within the OL community on the exact task they are concerned with (Buitelaar, Cimiano, & Magnini, 2005; Maimon & Browarnik, 2015; Wong et al., 2012). Following a closer look into many OL studies, it is clear that most existing OL systems suffer many drawbacks and shortcomings. There is a consensus among several aspects of ontology construction challenges that require more efforts. The following list presents the common aspects that define the main challenges of OL regarding concepts extraction and relations extraction methods:

i. Fully automatic construction for ontologies could not be possible. Still, there is an acute need for more effort to decrease human intervention in the ontology construction process to build (semi-) automatic construction rather than the existing cooperative systems of ontology construction (Buitelaar et al., 2005; Maimon & Browarnik, 2015; Mishra & Jain, 2015; Sathiya & Geetha, 2018; Siebra & Wac, 2023; Wong et al., 2012; Zhou, 2007).

ii. There is a need to avoid the noise terms (irrelevant or very general) that lead to unnecessary additional efforts. This issue could be addressed by paying more attention to filter terms in the ontology construction process as early as possible, i.e., in the early stage of construction (Gillani Andleeb, 2015; Konys, 2018; Wong et al., 2012).

iii. Extracting the concepts and the relations between them are still unsatisfactory in terms of its result, and more efforts are needed in this aspect. (Albukhitan, Helmy, & Alnazer, 2017; Arefyev et al., 2019; Buitelaar et al., 2005; Maimon & Browarnik, 2015; Wong et al., 2012; Zhou, 2007).

iv. The transformation of data from small to large data should be taken into account when designing an OL system (Buitelaar et al., 2005; Mishra & Jain, 2015; Wong et al., 2012; Zhou, 2007).

The validity of these aspects and challenges will be evident from further discussion on the literature and state-of-the-art of OL in Chapter 2.

1.2 Related Tasks of Ontology Learning

1.2.1 Concepts Extraction

Concepts extraction is the baseline task for an OL system. This task aims to extract relevant concepts (meaningful concepts that represent the domain knowledge) and filter out insignificant concepts. One of the concept extraction challenges is that there is no consensual or clear definition of what the formation of the concept is. (Buitelaar et al., 2005; Maimon & Browarnik, 2015; Petrucci, Ghidini, & Rospocher, 2016). In this study, relevant concepts can be defined as the terms, objects, and their instances that define and draw the outlines, keywords, and classes of knowledge within a domain. Following many OL studies such as (Gillani Andleeb, 2015; Völker, Fernandez Langa, & Sure, 2008), the concept formation in this research is defined as the following tuple: $< \boldsymbol{x}, \boldsymbol{L} >$ where $\boldsymbol{x}$ refers to the concept intension or sign, and $\boldsymbol{L}$ refers to the concept lexical intention (linguistic realization).

1.2.1(a) Concept Relevance Measurements

Concept Relevance Measurements are the metrics that are used to determine relevant concepts within a domain, such as Term Frequency-Inverse Document Frequency (TF-IDF) and C-value. Using relevance metrics is important to avoid the noise terms (irrelevant or very general) that affect the performance results and lead to unnecessary additional efforts (Chau, Labutov, Thaker, He, & Brusilovsky, 2021; Wong et al., 2012; Zhou, 2007). Addressing and filtering noise terms in the early stage of OL improves and increases the quality of ontology construction.

Many of the existing concept extraction methods are based on relevance metrics that determine relevant concepts for the target domain by estimating the relevancy for each concept within the target and contrasting domains simultaneously, such as (X. Jiang & Tan, 2010; YAO, GAN, & XU, 2017). Simultaneously processing the target and contrasting domains to identify the relevant concepts could be a high computation cost for large-length data.

The target domain term refers to the domain in which the dataset (in a specific field) is collected to construct the ontology within it. In contrast, a contrasting domain term refers to the public or general domain with a heterogeneous dataset collected from different fields. In addition, a large-length dataset means that the corpus contains a large number of sentences and words. In other words, it has a large Bag-of-Words (BOW); the size of the BOW is the number of the corpus words without redundancy. Likewise, a small-length dataset means that the corpus contains a small BOW size.

Indeed, the existing relevance metrics do not estimate the concepts sustainability distribution across the dataset. Sustainability distribution means estimating the distribution value of the concept if it is sustainable (continuous with approximately smooth distribution differences) through the corpus. In other words, it means the continuous appearance of concepts across the text. This thesis introduces a new strategy for estimating the sustainability distribution of concepts based on the time factor, by ordering the corpus from old to recent then estimating the concepts' distribution differences for time across the corpus (more explanation will be in Chapter 4). To explain the importance of sustainability distribution based on the time factor, suppose that at the beginning of 2020, we tried to construct ontology in the medical domain for fast-spreading diseases/viruses as a sub-domain (e.g. Influenza, MERS-CoV, SARS-CoV, and SARS-CoV-2 (COVID-19)). So using the existing metrics, COVID-19 would not be identified as a significant relevant concept because its frequency is less than others and because it appears in a few articles.

More details and elucidations regarding concepts extraction and relevance metrics will be presented in Chapter 2.

1.2.2 Relations Extraction

Relations refer to the connections or associations between different concepts. They are used to represent the relationships between concepts in a domain. Relations extraction, also called relations discovery, aims to discover and extract the taxonomic and non-taxonomic relationships among selected concepts. Taxonomic relations extraction aims to build the hierarchical taxonomy of concepts, while non-taxonomic relations extraction aims to extract the semantic relations among selected concepts. Discovering the relationships among the relevant concepts within a domain is considered the backbone of an OL system (Buitelaar et al., 2005; Maimon & Browarnik, 2015).

Relation extraction techniques are a combination of NLP and ML. Indeed, the relation extraction techniques could be fallen into four classes: (i) linguistic-based approaches that prioritize the use of syntactic analysis and patterns, (ii) statistical-based approaches that base on associations and hierarchical analysis and rules, (iii) logic-based approaches that explore logic theories in an attempt to infer or derive rules, and (iv) hybrid approaches that generally blend the traditional techniques, with the new approaches in current exist studies such as using artificial neural networks (ANN). Deep Learning (DL) is a type of ANN that has more than one hidden layer. Deep Belief Network (DBN) and Convolutional Neural Network (CNN) are examples of DL models. DBN will be used in this research.

More details and elucidations about relations extraction methods, approaches, defects, and obstacles as well as DL will be presented in Chapter 2.

1.3 Motivation

With the explosive growth of textual data on the Web, the motivation of representing the knowledge in this data in the machine-understandable form (ontologies) is increased. OL shaped a somewhat different vision of knowledge acquisition; it handles the knowledge acquisition bottleneck of actually shaping for the relevant knowledge to the interest domain. Many recent applications are based on ontologies for modelling and drawing the knowledge and results (e.g., IR, QA, and semantic searching).

Concepts and relations extraction are the principal tasks for any OL system and the most challenging tasks. The consistency and integrity of an ontology depend on the quality and precision of extracted concepts and relations (Sathiya & Geetha, 2018). The current OL systems have many shortcomings and drawbacks for extracting concepts and relations, such as using a small dataset, depending on a large amount of manually predefined patterns, and extracting very few types of relations. Hence, there is a need for effective and efficient techniques and models that aim to improve the concepts and relations extraction tasks for OL.

1.4 Problem Statement

The OL systems change the way of processing textual data from text mining to knowledge mining. This knowledge has to represent in the form of concepts and relationships between those concepts (ontologies) to be in machine-understandable form. Consequently, there is a serious need to design and develop modern systems and techniques that can represent the knowledge by automatically extracting and constructing ontologies.

Concepts and relations extraction are the backbone of an OL system. In other words, concepts and their relations are the prime components of an ontology. Through a closer look at the different OL systems and approaches, many shortcomings of current OL systems can be reviewed. One of the major shortcomings in the existing OL systems is that the existing concepts extraction methods (including relevant measurements) are not efficient to extract the relevant concepts and avoid noisy data (Gillani Andleeb, 2015; Konys, 2018; Wong et al., 2012). They are extracting relevant concepts from target and contrasting domains without taking into account the transformation to large-length data, such as in (Chau et al., 2021; X. Jiang & Tan, 2010; M. K. Kim, Zouaq, & Kim, 2016; A Maedche & Volz, 2001; YAO et al., 2017). Estimating the relevancy value of concepts in target and contrasting domains causes a high computational cost. Also, using the contrasting domain to identify the noise terms (irrelevant or very general) synchronizing with concepts relevance estimating causes unnecessary additional efforts, and may affect the quality and precision of results. In addition, the hypothesis that if this concept is frequent in a specific field, then it becomes insignificant in the other field is not always correct. For example, *image processing* is a significant concept in the *deep learning* field and in the *computer vision* field, so its repetition in another field does not necessarily mean that this concept is not significant in this field. Furthermore, it doesn't make sense to collect contrasting data approximately equivalent to the data of the target domain otherwise, the metric cannot be able to estimate the relevancy, especially for large data. As well, these existing studies cannot extract the modern relevant concepts within a domain because they do not estimate the concepts sustainability distribution across the dataset.

Therefore, there is a need to develop efficient automatic methods for improving concepts extraction task for OL from only the target domain, taking into account the transformation of data from small to large-length data with avoiding the noise terms at early stages. Furthermore, the proposed methods should be capable of extracting modern relevant concepts.

Another significant shortcoming in existing OL systems is their failure to adequately extract semantic relations. These systems typically extract a limited array of relations (most of them are taxonomic), almost do not exceed synonyms, hyponyms, hypernyms, meronyms, and holonyms relations such as in (Arnold & Rahm, 2014; Doan, Arch-int, & Arch-int, 2020; Gillani Andleeb, 2015; Lung-Hao Lee, 2017; Qiu, Qi, Wang, & Zhang, 2018). And most of them are based on a large amount of manually predefined patterns to extract relations. Manually predefined patterns have reasonable precision but a very low recall (Arefyev et al., 2019; Gillani Andleeb, 2015; Maimon & Browarnik, 2015; Sathiya & Geetha, 2018; Wong et al., 2012). Additionally, these patterns are frequently restricted by the scope and coverage of the rules used and can be challenging to adapt to new domains. Moreover, many existing studies extract any verb according to their predefined patterns and designate them as relations, disregarding the importance of relations categorization, such as (Cimiano & Völker, 2005; X. Jiang & Tan, 2010; Saber, Abdel-Galil, & El-Fatah Belal, 2022; Völker et al., 2008).

Recently, some studies, such as (Ben Abdessalem Karaa, Alkhammash, & Bchir, 2021; Bergelid, 2018), have turned to machine learning (ML) algorithms, like the Support Vector Machine (SVM), as a potential solution. These methods may be more flexible; however, they do not perform optimally with large or noisy data and often require a substantial amount of clean data for training. In essence, existing relation extraction methods struggle with shallow text analysis and comprehension. Therefore, there is a pressing need to develop adaptive relation extraction methods for

OL that can perform deep analysis, adapt to new domains and noisy or large data, and extract semantic relations automatically.

1.5 Research Questions

To address the problems stated in the problem statements section, this thesis focuses on answering the following questions:

i. How to enhance a concepts extraction method (including proper relevance metric) for effective extraction of relevant and modern concepts, taking into account the transformation to large data?

ii. How to design an adaptive relations extraction method for OL that can efficiently extract and classify semantic relations from noisy or large data?

1.6 Research Objectives

The fundamental goal of this research is to define, develop and create a new adaptable framework for automatically extracting relevant concepts and semantic relations for OL, with estimating the concept sustainability distribution across long-length domain datasets and with more efficient extraction methods. In other words, this research has two objectives:

i. To propose a method for concepts extraction from scientific publications (with a proper relevance metric) for handling long-length datasets and avoiding the noise.

ii. To propose an adaptive relations extraction method for OL that can adapt to new domains with noisy or large data, as well as can extract and classify semantic relations.

1.7 Research Contributions

This research aims to address the research gaps in relevant concept and relation extraction tasks on the OL with taking into account the large-length data and avoiding the noise data in the early stage of the OL process. The main contribution of this research is to create and develop a new Concept and Relation Extraction Framework (CREF) that is adaptable for domain ontology construction across the long-length dataset, with a high F1 score. These main contributions could be pointed out as the following:

i. A Developed Concepts Extraction Method, called (DTR-DCEM), based on new Domain Time Relevance metric (DTR) for sustainability distribution estimating and using the proposed Concepts Extraction stopwords (CE-stopwords) for scientific publications to improve relevant and modern concepts extraction from target domain only and can handle large-length data with avoiding noise terms.

ii. An adaption of the DBN model called Re-DDBN coupled with a new text representation approach called Binary Concept-Verb Bag-of-Words (BCV-BOW) to improve the relations extraction task for OL with ability to adapt with new domains and handle noisy and large data to extract and classify semantic relations.

1.8 Research Scope

The area of this research is OL for creating a new ontology. This thesis focuses on concepts and relations extraction tasks of the OL process. As well, this research has been conducted in the scientific publications domain for the English language.

This research studies the different OL systems and approaches for automatic concepts and relations extraction. OL process consists of six layers, namely, *terms layer*, *synonyms layer*, *concepts layer*, *concepts hierarchies layer*, *relations layer*, and *rules layer*. This research deals with the three middle layers, they are *concepts layer*, *concepts hierarchies layer*, and *relations layer*. So, this research will not deal with the rest layers of the OL process, which are the first, second, and last layers.

In short, this research focuses on creating and developing the best methodology for the relevant concepts extraction task for OL. In addition, this research also focuses on improving the semantic relations extraction task for the OL process. Moreover, this research deals with how to use and address DBN to improve relations extraction task for OL.

1.9 Thesis Outlines

This chapter has presented the background of ontology construction, as well as it has explained the research problem statement, motivation, questions, objectives, and contributions. The rest chapters of this thesis are organized as follows:

Chapter 2: presents the literature review, which includes recent studies regarding OL systems, tasks, and approaches. Also, it explores the related works about concepts extraction and relations discovery for OL systems. In addition, it presents the DL overview and text representation approaches for ML and DL.

Chapter 3: elaborates the methodology and the proposed framework for designing the CREF for OL. In addition, it describes the datasets and benchmarks, and defines the main performance metrics for system evaluation.

Chapter 4: presents the proposed enhancement of the concept extraction and formation method called DTR-DCEM using the proposed DTR and CE-Stopwords. And it shows the comparison between the proposed method and the traditional methods.

Chapter 5: elaborates the traditional DBN, its configuration as well as its structure, its hyper-parameters, its functions, and its algorithm. Also, it presents the proposed approach for text representation for DL called BCV-BOW. It also shows the imbalance of the datasets, the training results for relation extraction using the Re-DDBN model, and the comparison of the proposed approach and method with the benchmark models.

Chapter 6: presents and elaborates the overall analysis of all the results of the proposed CREF framework for constructing ontologies. Also, it outlines the conflicts in concepts extraction and presents the relations extraction and classification results using the trained Re-DDBN.

Chapter 7: concludes the thesis, explains the research limitations and highlights several other possible enhancements for future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Ontology Learning Overview

Ontology plays an important role in Knowledge Management and the Semantic Web. OL from text is a process that aims to automatically represent the knowledge in machine-understandable form. In general, ontologies represent the knowledge related to a specific domain in terms of the relevant concepts and the relationships between these concepts to be in machine-readable form. In this section, the OL systems and approaches with their criticism are presented.

2.1.1 Ontology Construction via Learning

OL is a process of automatically or semi-automatically creating a new ontology or reusing (for enriching or populating) an existing ontology; with the minimum exertion of a human (Gillani Andleeb, 2015). OL from text is a process of acquiring knowledge from a text by applying a set of methods and techniques from various fields, such as NLP and ML, for extracting ontological elements (Maimon & Browarnik, 2015) and then constructing ontologies.

Ontology Learning Layer Cake was proposed by Buitelaar et al. (2005) as in Figure 2.3. This approach is a dominant approach, and it is considered the cornerstone in OL (Maimon & Browarnik, 2015). According to Buitelaar et al. (2005), there are six layers in OL; they are *Terms, Synonyms, Concepts, Concepts Hierarchies, Relations,* and *Rules.* Based on these layers, the process of OL can be divided into five sub-tasks as follows:

i. *Terms Extraction* is a prerequisite for all aspects of OL from the text. A term is a multi-word or single-word token, which denotes a specific meaning in a given domain.

ii. *Synonyms Discovery* aims to find the terms that indicate the same concept and appear in the same set for a selected concept. A synonym set is the most common type of lexical relations. It is possible to use a readily available set such as WordNet *synsets*, clustering techniques, or any other similar methods (Maimon & Browarnik, 2015).

iii. *Concepts Extraction* is also called concept formation*;* it is a baseline task for OL. But this task is considered unclear because there is no consensual definition of what a concept is (Buitelaar et al., 2005; Maimon & Browarnik, 2015; Petrucci et al., 2016). A concept, in general terms, is defined by two main parts: the intension signs which encompass the label or characteristics that uniquely identify the concept, and the lexical intention which describes the words or phrases in language to point to the concept.

iv. *Relations Extraction* is the backbone of an OL; it aims to extract the taxonomic and non-taxonomic relations. Taxonomic relations extraction, also called concept hierarchy, aims to build the hierarchical taxonomy of concepts (is_A and has_A relations). Non-taxonomic relations extraction, also called semantic relations extraction, aims to extract novel semantic relationships between known concepts. There are a few approaches that address the relation extraction issue for OL; this task is still an open problem (Gillani Andleeb, 2015; Maimon & Browarnik, 2015; Mishra & Jain, 2015; Sathiya & Geetha, 2018; Wong et al., 2012).

v. *Rules or axioms extraction* is the final sub-task in the OL process. It aims to infer the rules based on extracted concepts and relations. This task is based on the relations extraction, and it is still in the initial stage and needs more efforts (Gillani Andleeb, 2015; Maimon & Browarnik, 2015; Mishra & Jain, 2015; Wong et al., 2012; Zhou, 2007). There are few attempts to generate the rules and axioms in existing OL systems. So far, logic-based approaches can be used for this task such as in (Fleischhacker & Völker, 2011; Oliveira, Pereira, & Cardoso, 2001).

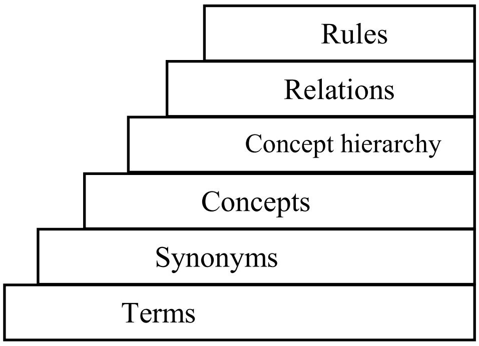
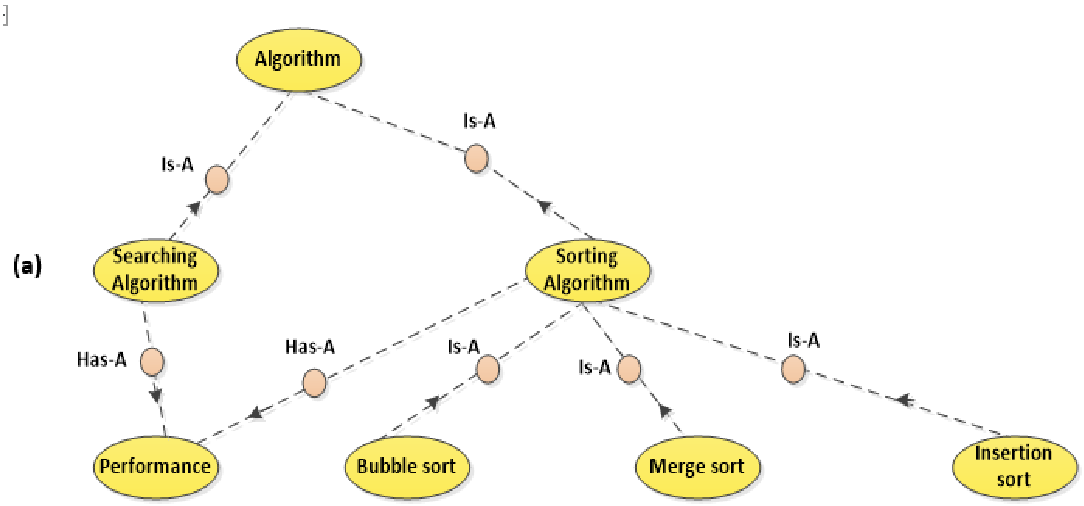


Figure 2.1 Ontology Learning Layer Cake

In short, we can define the OL process as the automatic identifying the relevant concepts and their relationships, mapping and representing the ontology by representation language (e.g., OWL or RDF). Figure 2.2(a) shows a graph example of ontology in "*Data Structures*" domain, while Figure 2.2(b) shows the annotation of this example that can be encoded by ontology representation languages.



```
(b) (Domain, Data structure)                          // domain
    (Class,  Algorithm)                               // c1
    (SubClass, Sorting algorithm, Algorithm)          //(r1,  c2, c1)
    (SubClass, Searching algorithm, Algorithm)        //(r1,  c3, c1)
    (Has/Property, Performance, Sorting algorithm)    //(r2,  c4, c2)
    (Has/Property, Performance, Searching algorithm)  //(r2,  c5, c3)
    (SubClass, Bubble sort, Sorting algorithm)        //(r1,  c6, c2)
    (SubClass, Merge sort, Sorting algorithm)         //(r1,  c7, c2)
    (SubClass, Insertion sort, Sorting algorithm)     //(r1,  c8, c2)
```

c_i: concept r_i: relation

Figure 2.2 An example of Ontology

2.1.2 Approaches on Ontology Learning from Text

OL approaches can be divided into linguistics-based approaches, ML approaches (statistic-based and logic-based approaches) and hybrid approaches. The following sub-sections show the most well-known approaches on OL from the text:

2.1.2(a) Linguistics-Based Approaches

i. *Pattern-based extraction* (Morin, 1999): It is used to recognize the relations by matching a pattern from a sequence of words in the text. Lexico-syntactic patterns and semantic templates are techniques under this approach. The lexico-syntactic patterns technique uses any defined patterns such as “NP is a type of NP" to extract hypernym and meronym relations. Semantic templates are similar to the lexico-syntactic patterns technique but with more detailed rules and conditions and also has been used to extract non-taxonomic relations. It is well known that these approaches have reasonable precision, but they have very low recall (Maimon & Browarnik, 2015; Wong et al., 2012).

ii. *POS tagging and sentence parsing* (Abney, 1997): It is considered a rule-based approach. Part-Of-Speech (POS) tagging is used to assign parts of speech to each word in the text, such as noun, verb, adjective, etc., while the sentence parser is used to recover the complete and exact parses for each sentence in the text. However, many words are ambiguous (e.g., in English, the word "plant" may be a noun or a verb), so certain parsers are built on statistical-based parsing, such as the Stanford Parser (Klein & Manning, 2003). Statistical Parsing is based on the probability of certain tags occurring, given various possibilities. It uses the probability of certain sequences of tags occurring to determine the most likely syntactic structure of the sentence. This approach is used for term extraction.

iii. *Syntactic structure analysis and dependency structure analysis* (Gamallo, Gonzalez, Agustini, Lopes, & De Lima, 2002; Nivre, 2004): It is used to uncover the syntactic and dependency of terms and relations at the sentence level. Syntactic structure analyses the words and modifiers in syntactic structures such as noun phrases and verb phrases to discover potential terms and relations while ignoring other phrases. As for dependency structure analysis, it uses grammatical relations (e.g., subject, object, and complement) to determine more complex relations. However, these approaches on their own may be inadequate; they need to interact and complement with other algorithms and rules for improved performance (Mudhsh, Al-Takhayinh, & Al-Dala'ien, 2015). In other words, these techniques may not accurately identify every concept or relation, for example, it cannot be asserted that all noun phrases are concepts. Hence, they must work in tandem with other techniques for more accurate and comprehensive concept identification. This approach is useful for term and concept extraction, and also for relation extraction. For example, concepts can be extracted based on terms dependency within a noun phrase, while relations can be extracted based on terms dependency within a verb phrase.

2.1.2(b) Machine Learning Approaches

ML approaches can be divided into two kinds: statistical-based approaches and logic-based approaches.

2.1.2(b)(i) Statistic-Based Approaches

i. *Co-occurrence analysis* (Budanitsky, 1999)*:* It is used to identify lexical units that tend to occur together for purposes ranging from extracting related terms to discovering implicit relations between concepts.

ii. *Association rules* (Alexander Maedche & Staab, 2000): It is used to extract the non-taxonomic relations between concepts by using a small seed knowledge as background (e.g., using concept hierarchy as background).

iii. *Heuristic and conceptual clustering* (Faure & Nédellec, 1998; Faure & Poibeau, 2000): It is used to group the concepts based on the semantic distance between them to make up hierarchies. Formal Concept Analysis (FCA) is one method under this approach that uses the conceptual clustering technique to provide intentional descriptions for the abstract concepts or data units (Cimiano, Hotho, & Staab, 2005; Drymonas, Zervanou, & Petrakis, 2010).

iv. *Ontology pruning* (Kietz, Maedche, & Volz, 2000): It is used to build a domain relevant ontology by using heterogeneous sources (e.g., using relevant metrics or comparing domain sources with the generic sources to determine which concepts are more relevant to the specific domain and which concepts are general).

2.1.2(b)(ii) Logic-Based Approaches

i. *Inductive Logic Programming* (Zelle & Mooney, 1993): It is used to derive the rules from positive and negative examples of the existing collection of concepts and relations. For example, firstly, the following positive examples: "cats have fur", "dogs have fur", and "tigers have fur", then "mammals have fur" are generated. After that, from the negative example, "humans do not have fur", then the generalization of "mammals have fur" will be dropped and deduced that only "canines and felines have fur". However, this approach depends on the good predefined rules templates by the expert. For instance, if there are no good negation examples, then an invalid rule or fact may be generated. The considerable disadvantage of this approach is that in the search process, it sometimes can prune searched hypothesis (Boytcheva, 2002).

ii. *Logical inference* (Shamsfard & Barforoush, 2004): It is used to infer implicit relations from existing ones. For example, "Steven is a man" and "all men are mortal", then the following relation "Steven is mortal" is inferred. However, in this approach, there is a high possibility of introducing conflicting and/or invalid relations and rules (Wong et al., 2012). For example, “human eats fish” and “fish eats the worms” potentially generate invalid new relation. In addition, it can generate only very basic relations most of the time.

2.1.2(c) Hybrid Approaches

Hybrid approaches strive to merge traditional techniques with advanced methods found in current research. They typically involve combining linguistic, and statistical-based methodologies with other ML algorithms such as Support Vector Machines (SVM), Naive Bayes (NB), Artificial Neural Networks (ANN), or DL models like Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Deep Belief Networks (DBN). Hybrid approaches could typically be categorized into three sub-approaches as follows:

i. *Machine Learning Algorithms Combined with Linguistic Features:* It utilizes ML techniques paired with carefully chosen linguistic features to extract more nuanced information from text data such as in (Cambria, Poria, Bisio, Bajpai, & Chaturvedi, 2015; Lung-Hao Lee, 2017). It is particularly efficient at tasks like terms extraction, and relations extraction, where both statistical patterns and linguistic insights can be leveraged for improved performance. The major disadvantage of this approach is its high dependency on the quality and specificity of the linguistic features used (Mehanna & Mahmuddin, 2021).

ii. *Deep Learning Models Integrated with Linguistic Analysis:* These models can capture more complex relationships and dependencies within the data. They are especially potent at tasks like concepts and relations extraction, owing to their capability to handle ambiguity, variability, and complexity in text. However, the main drawback of these approaches lies in their high computational cost, and they might require large amounts of data to work efficiently (Goodfellow, Bengio, & Courville, 2016).

iii. *Machine Learning or Deep Learning Integrated with Statistical Techniques:* These approaches harness the strengths of ML or DL models and integrate them with statistical techniques to identify patterns, correlations, and frequency distributions in text data, supporting tasks such as terms extraction, synonyms discovery, and concepts extraction. However, the performance of these models is heavily dependent on the quality of the statistical analysis, as low-quality or insufficiently nuanced statistical analysis can limit the efficacy of the model (Gupta & Lehal, 2009).

Concluding this approaches section, Table 2.1 shows the summary of the discussed OL approaches and OL tasks corresponding with these approaches.

Table 2.1 Ontology Learning Approaches and their Corresponding Tasks

OL Approaches		OL Tasks	OL Approaches Disadvantages	OL Approaches Advantages
Linguistics-Based Approaches	*Pattern-based extraction:* – *Lexico-syntactic patterns* – *Semantic templates*	Relations extraction	Have very low recall	Has reasonable precision
	POS Tagging	Terms extraction	Ambiguity (one word may have more than one tag)	Good for characterizing the context
	Statistical parsing	Terms &Concepts extraction	Ambiguity (but less than POS)	Good for uncovering the syntactic and grammatical relations in context

Table 2.1 Ontology Learning Approaches and their Corresponding Tasks

OL Approaches		OL Tasks	OL Approaches Disadvantages	OL Approaches Advantages
	Syntactic structure analysis& dependency structure analysis	Terms, Concepts & Relations extraction	Need to corporate with other algorithms or rules to have better performance	Has good results for extracting related terms that occur together
Statistic-Based Approaches	*Co-occurrence analysis*	Terms extraction & Concepts extraction	Not appropriate for relation extraction task.	Has good results for extracting related terms that occur together
	Association rules	Relations extraction	Need large support factor specified in advance	Has good results for well-defined problems
	Heuristic/ conceptual clustering	Synonyms discovery, Concepts extraction & Taxonomic relations extraction	Not applicable for non-taxonomic relation extraction	Good for grouping the concepts based on the semantic distance
	Ontology pruning	Terms extraction	Not applicable for relation extraction task	Good for reducing the noise data by determining the relevant concepts
Logic-Based Approaches	*Inductive Logic Programing*	Axioms extraction	May introduce invalid rules or facts& May prune the searched hypothesis	Has good results for the good predefined rules problems
	Logical inference	Relations extraction	May infer invalid or conflicting relations & may only generate very basic relations	Can generate new basic rules or facts from existing ones