

CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD



Citations' Context and Reasons
Ontology - CCRO: A Model for
Citation Reasons between
Research Articles

by

Imran Ihsan

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the

Faculty of Computing

Department of Computer Science

2021

**Citations' Context and Reasons Ontology -
CCRO: A Model for Citation Reasons between
Research Articles**

By

Imran Ihsan

(DCS 143007)

**Dr Hélène de Ribaupierre, Lecturer
Cardiff University, Cardiff, UK
(Foreign Evaluator 1)**

**Dr. Klaus Tochtermann, Professor
Kiel University, Kiel, Germany
(Foreign Evaluator 2)**

**Dr. Muhammad Abdul Qadir
(Thesis Supervisor)**

**Dr. Nayyer Masood
(Head, Department of Computer Science)**

**Dr. Muhammad Abdul Qadir
(Dean, Faculty of Computing)**

**DEPARTMENT OF COMPUTER SCIENCE
CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
ISLAMABAD**

2021

Copyright © 2021 by Imran Ihsan

All rights reserved. No part of this thesis may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.

To My Father



**CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY
ISLAMABAD**

Expressway, Kahuta Road, Zone-V, Islamabad
Phone: +92-51-111-555-666 Fax: +92-51-4486705
Email: info@cust.edu.pk Website: <https://www.cust.edu.pk>

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the thesis, entitled “**Citations' Context and Reasons Ontology - CCRO: A Model for Citation Reasons between Research Articles**” was conducted under the supervision of **Dr. Muhammad Abdul Qadir**. No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the **Department of Computer Science, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Computer Science**. The open defence of the thesis was conducted on **February 09, 2021**.

Student Name : Imran Ihsan (DCS143007)

The Examination Committee unanimously agrees to award PhD degree in the mentioned field.

Examination Committee :

(a) External Examiner 1: Dr. Sharifullah Khan,
Professor
PAF-IAST, Haripur, KPK

(b) External Examiner 2: Dr. Hammad Majeed,
Associate Professor
FAST-NUCES, Islamabad

(c) Internal Examiner : Dr. Azhar Mahmood
Associate Professor
CUST, Islamabad

Supervisor Name : Dr. Muhammad Abdul Qadir
Professor
CUST, Islamabad

Name of HoD : Dr. Nayyer Masood
Professor
CUST, Islamabad

Name of Dean : Dr. Muhammad Abdul Qadir
Professor
CUST, Islamabad

AUTHOR'S DECLARATION

I, **Imran Ihsan** (Registration No. DCS143007), hereby state that my PhD thesis titled, '**Citations' Context and Reasons Ontology - CCRO: A Model for Citation Reasons between Research Articles**' is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my PhD Degree.



(**Imran Ihsan**)

Dated: 9 February, 2021


Registration No : DCS 143007

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled “**Citations' Context and Reasons Ontology - CCRO: A Model for Citation Reasons between Research Articles**” is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/ cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of PhD Degree, the University reserves the right to withdraw/ revoke my PhD degree and that HEC and the University have the right to publish my name on the HEC/ University Website on which names of students are placed who submitted plagiarized thesis.



(**Imran Ihsan**)

Dated: 9 February, 2021

Registration No : DCS 143007

List of Publications

It is certified that following publication(s) have been made out of the research work that has been carried out for this thesis:-

1. **I. Ihsan**, and M. A. Qadir, "CCRO: Citations' Context and Reasons Ontology," *IEEE Access*, vol. 7, no. 1, pp. 30423-30436, 2019.

Imran Ihsan

(DCS 143007)

Acknowledgements

While writing this acknowledgement page, I am remembering those days when I started working on my research thesis. I started from nothing and now I am ending up with a lot of useful discussions, information and achievements. At this moment of glory, I must thank to all mighty ALLAH, that he has given me strength and courage to complete this work.

As I imagine is the case with most authors, my family comes next on the list of those whose help and support I gratefully acknowledge. Time spent in completing the work usually meant time “stolen” from other worthy activities. I do not recall one serious complaint from my Wife, Son and Daughters.

Next on the list would have to be my supervisor Dr. Muhammad Abdul Qadir. I am thankful for there comments, academic support, direction and guidance that helped to give some sense of structure to the work.

I owe many thanks to my friends and colleagues at Air University for their discussions, continuous moral support and prayers. I could not have finished it without their help.

Imran Ihsan

Abstract

Research papers can be visualized as a networked information space that contains a collection of information entities, inter-connected by directed links, commonly known as Citation Graph. There is a possibility to enrich the citation graph with meaningful relations between the citing and the cited articles to express the citation's reason using semantic tags. We have explored the existing tags and evaluated them against the representation of the citation's context and reasons. We have discovered more than 150 citations' reasons from the published literature to be represented as citation tags. Many of these reasons have overlapped and diffused meanings. A citation graph is a forest of graphs with hundreds of nodes in each graph. Annotating such a large volume of graphs with citation's reasons manually, requires a huge effort, and is nearly impossible. Thus, giving rise to a need to discover the citation's reasons automatically with a high accuracy. In order to achieve this, the first step is to develop a minimal set of citation's context and reasons that are disjoint in nature (if possible). It would be great help to the reasoning system if these reasons are represented in a formal way in the form of Ontology. A formally defined set of reasons can make machine-learning algorithms to identify these reasons. By adopting a well-defined scientific methodology to formulate an ontology of citation reasons, we have reduced 150 reasons into only eight reason classes by using an iterative process of sentiment analysis, collaborative meanings and experts' opinions. Based on our findings and experiments, we have proposed an *Ontology for Citations' Context and Reasons – CCRO* that provides abstract conceptualization required to organize citations' relations. *CCRO* has been verified, validated and assessed by using well-defined procedures and tools proposed in the literature for ontology evaluation. The results show that the proposed ontology is concise, complete and consistent. For the instantiation and mapping of ontology classes on real data, we have developed a *Mapping Graph* between the verbs with predicative complements in English Language, the verbs extracted from the selected corpus using *NLP* and our *CCRO* classes.

In recent scientific advances, Artificial Intelligence and Natural Language Processing are the major contributors in classifying documents and extracting information. Classifying citations in different classes has gathered a lot of attention due to large volume of citations available on different digital libraries. Typical citation classification is based on sentiment analysis, where various techniques are applied on citations texts to mainly classify them in “Positive”, “Negative” and “Neutral” sentiments. Using *CCRO*, next step adapts an ontology-based approach to extract citation’s reasons and instantiate ontology classes and properties on two different corpora of citation sentences. One corpus of citation sentences is a publicly available data-set, while the other is our own manually curated. The process uses a two-step approach. First part is an interface to manually annotate each citation text in the selected corpora on *CCRO* properties. A team of carefully selected annotators have annotated each citation to achieve high inter-annotator agreement. Second part focuses on automatic extraction of these reasons. Using *Natural Language Processing, Mapping Graph* and *Reporting Verb* in a citation sentence, citation’s reason is extracted and mapped onto a *CCRO* property. After comparing both manual and automatic mapping, accuracy is calculated. Based on experiments and results, our algorithm shows overall accuracy of 85.4% and 96.6% in publicly available and our own corpora of citation sentences respectively.

The number of research articles in today’s world has grown exponentially. With such a huge digital infrastructure, where there is a need to gather actionable intelligence from millions of papers that are without any useful semantics, there is also a need to improve the ways where new research articles are authored and disseminated to build knowledge base. In order to look at both sides the problem, two different application of *CCRO* are formulated. One that deals with the legacy data and the other that deals with the authoring of new research with useful semantics.

A citation graph has the potential to reveal important and interesting information about the history of a particular scholarly research that has happened during its life-cycle. Citation graphs can be enriched with semantic tags, where scientific papers are inter-connected with citation reasons. For the first application, using

CCRO properties, our selected corpora are initially converted to *Semantic Citation Graph*. With help of guidelines provided in literature, five different queries are then formulated to discover evolutionary path of a scholarly activity, to find current state of a research, and to examine different school of thoughts around a problem etc.

However, one of the best source of knowledge to tell the reason of citation is the author of the paper. Authors of the scholarly articles cite other articles, based on certain reasons. Integrating these citations' reasons in an authoring system can help authors to choose a reason while citing. So for our second application, a **Semantic L^AT_EX**, that integrates *CCRO* properties within L^AT_EX document, is proposed. Using *CCRO* properties, to semantically tag citations with reasons can create an discourse relation between research papers. Furthermore, embedding these structures within *RDF* Data Store enables the creation of semantic publications that becomes a foundation artifact for the Semantic Publishing Ecosystem and linked resources become part of the current Web of Data.

Contents

Author’s Declaration	v
Plagiarism Undertaking	vi
List of Publications	vii
Acknowledgment	viii
Abstract	ix
List of Figures	xvi
List of Tables	xviii
Abbreviations	xx
Symbols	xxi
1 Introduction	1
1.1 Citation and Citation Graph	1
1.2 Citation Analysis	2
1.3 Citation Context and Reasons	3
1.4 Open Citations	5
1.5 Scientific Authoring of Citation Reasons	6
1.6 Motivation	6
1.7 Problem Statement	7
1.8 Research Questions	8
2 Literature Review	10
2.1 Semantic Representations	11
2.1.1 BiRO - Bibliographic Reference Ontology	11
2.1.2 C4O – Citation Counting and Context Characterization Ontology	11
2.1.3 FaBio – FRBR-aligned Bibliographic Ontology	12
2.1.4 DoCO – Document Component Ontology	12

2.1.5	SWAN – Discourse Ontology	12
2.1.6	PRO – Publishing Role Ontology	13
2.1.7	PSO – Publishing Status Ontology	13
2.1.8	CiTO – Citation Typing Ontology	13
2.2	CiTO Limitations	14
2.2.1	Less Used Properties	14
2.2.2	Most Used Neutral Properties	14
2.2.3	Lower Inter-Rater Agreement	14
2.2.4	Non-Taxonomic Organization of CiTO-Ps	15
2.2.5	Customized Properties	15
2.2.6	Misinterpretation of Properties	15
2.2.7	Properties Perspective	15
2.3	Citation Reasons Classification	16
2.4	Semantic Authoring	22
2.5	Authoring Tools for Citation Reasons	22
2.5.1	Human Authoring and Annotation	22
2.5.2	Automatic Authoring and Annotation	24
2.6	L ^A T _E X Cite Packages for Citations	26
2.6.1	Cite Package	26
2.6.2	Harvard Package	26
2.6.3	Achicago Package	26
2.6.4	Natbib Package	27
2.6.5	Apacite Package	27
2.7	Conclusion	28
3	Methodology	30
3.1	Ontology Development	31
3.1.1	Primary Data - Citation’s Context and Reasons	33
3.1.2	Qualitative Data Analysis - Cluster Citation Reasons with Titles	33
3.1.3	Ontology Evaluation	33
3.2	Ontology Instantiation and Mapping	34
3.2.1	Secondary Data - Citations Corpora	34
3.2.1.1	Citation Corpus CC1:	34
3.2.1.2	Citation Corpus CC2:	35
3.2.2	Quantitative Data Analysis	36
3.2.2.1	Pre-Process Corpora	36
3.2.2.2	Manual Annotation of Citation Reasons	38
3.2.2.3	Automatic Mapping of Citation Reasons	38
3.2.2.4	Results and Findings	39
3.3	Ontology Applications	39
3.3.1	Query Semantic Graph	40
3.3.2	Semantic Authoring	41
3.4	Conclusion	42

4	Ontology Development	43
4.1	Knowledge Acquisition: Identify Key Terms and Concepts	44
4.1.1	Citation Context and Reasons	44
4.2	Ontology Conceptualization: Identify Classes and Properties	50
4.2.1	Cluster Citation Reasons	50
4.3	Formal Definition of Classes and Properties	52
4.3.1	CCRO Classes	52
4.3.2	CCRO Object Properties	54
4.3.3	CCRO Instance Level Schema	55
4.3.4	CCRO Citation Graph	56
4.4	Ontology Evaluation and Validation	57
4.4.1	Automated Tool Evaluation	57
4.4.2	User-Based Study Evaluation	58
4.4.3	Findings and Results	60
4.5	Conclusion	61
5	Ontology Instantiation and Mapping	63
5.1	Pre-Process Corpora	64
5.1.1	Extract Unique Verbs	64
5.1.1.1	Distribution of Verbs in Positive Sentiment	69
5.1.1.2	Distribution of Verbs in Negative Sentiment	69
5.1.1.3	Distribution of Verbs in Neutral Sentiment	69
5.1.2	Extract Reporting Verb	69
5.1.3	Generate Mapping Graph	70
5.2	Automatic Mapping of Citations' Reasons	72
5.2.1	Mapping Citation with Positive Sentiment on CCRO	74
5.2.2	Mapping Citation with Negative Sentiment on CCRO	74
5.2.3	Mapping Citation with Neutral Sentiment on CCRO	75
5.3	Manual Annotation of Citations' Reasons	76
5.4	Results	79
5.5	Findings	79
5.5.1	Combined Results Based on Sentiment	80
5.5.2	Combined Results Based on CCRO	80
5.5.3	Limitations in Automatic Mapping	81
5.6	Generate Semantic Knowledge Graph	82
5.7	Visualize Results	82
5.8	Conclusion	84
6	CCRO Application: Query Semantic Graph	87
6.1	Semantic Queries	88
6.1.1	What is the current state of debate on this question?	88
6.1.2	Which theory author disagrees with?	91
6.1.3	What assumptions does this approach depends on?	93
6.1.4	Are there different school of thoughts around this problem?	94

6.1.5	How many researches that have negated some other work? . . .	96
6.2	Conclusion	97
7	CCRO Application: Semantic Authoring	98
7.1	Methodology	100
7.1.1	Semantic Annotation	100
7.1.2	Semantic Authoring	101
7.1.3	Semantic Publishing	102
7.1.4	Semantic Graph	103
7.2	Experiment and Results	103
7.2.1	The CCRO Package	103
7.2.2	Semantic L ^A T _E X	105
7.2.3	Semantic Publishing Process	105
7.2.4	Semantic Citation Graph	106
7.3	Conclusion	108
8	Conclusion and Future Work	110
8.1	Conclusion	110
8.2	Research Contributions	112
8.3	Future Work	113
	Bibliography	115
	Appendix A:	
	Citation Reasons	129
	Appendix B:	
	Reporting Verbs	133
	Appendix C:	
	Citations' Context and Reasons Ontology	135
	Appendix D:	
	Questionnaire for User-based Evaluation of CCRO	142
	Appendix E:	
	CCRO Package	144
	Appendix F:	
	Existing Ontologies	147

List of Figures

1.1	Citation’s Context/Sentiment Examples	4
3.1	Methodology for Ontology Development and Mapping.	32
3.2	RDF Triple in RDF Data Store	40
4.1	Ontology Development Process.	43
4.2	Citation Reason Clusters and Properties.	51
4.3	Citation’s Context and Reasons Ontology	53
4.4	CCRO - Instance Level Schema Definition	56
4.5	CCRO - Citation Link RDF Triple Graph	57
4.6	Fact++ and Hermit Reasoners for CCRO	58
4.7	User Study Evaluation Report for CCRO	59
4.8	CCRO Properties Based on User-Based Study Evaluation	61
4.9	Example of “Negative” and “Neutral” Sentiment Citations for Same Citing and Cited Papers	62
5.1	Methodology for CCRO Instantiation and Mapping	64
5.2	Unique Verbs with Frequency ≥ 200 in AAN Corpus	65
5.3	High Occurrence Verbs Distribution in 3 Sentiments	66
5.4	Extract Reporting Verb from Citation Text using NLP and Verb Corpus	70
5.5	Partial Representation of Dictionary – A Mapping between Levin Conceptual Hierarchy Verbs and Citation Reason Classes	71
5.6	Mapping Model to Map Reporting Verb on CCRO Property using Mapping Graph and Sentiment Value	73
5.7	Mapping Citation with Positive Sentiment on CCRO	73
5.8	Mapping Citation with Negative Sentiment on CCRO	74
5.9	Mapping Citation with Neutral Sentiment on CCRO	75
5.10	CCRO Annotation Tool Interface	77
5.11	CCRO Instance for a Citing Paper - HI1001	77
5.12	CCRO Properties Distribution (Corpus CC1 - 7390 Citations)	77
5.13	CCRO Properties Distribution (Corpus CC2 - 1230 Citations)	77
5.14	Tabulated Results for Both Corpora CC1 and CC2	80
5.15	Combined Results on Sentiment for Both Corpora CC1 and CC2	80
5.16	Combined Results on CCRO for Both Corpora CC1 and CC2	81
5.17	Sample RDF File	83

5.18	Ontology Based Semantic Citation Graph for Corpus CC1 - Partial Visualization	84
5.19	Ontology Based Semantic Citation Graph for Corpus CC2 - H-Index Domain	85
6.1	Citation Sub Graph for Query 1 using Citation Corpus SCC2	89
6.2	SPARQL Query 1: Results Visualization	90
6.3	Citation Sub Graph for Query 2 using Citation Corpus SCC2	92
6.4	Citation Sub Graph for Query 3 using Citation Corpus SCC2	93
6.5	Citation Sub Graph for Query 4 using Citation Corpus SCC2	96
7.1	Steps towards Semantic Authoring and Publishing	101
7.2	RDF Triple in RDF Data Store	102
7.3	CCRO Package for L ^A T _E X- Syntax and Output	105
7.4	A Semantic L ^A T _E X Sample	106
7.5	A Semantic Publishing Process	107
7.6	CCRO Based Semantic Citation Graph	108

List of Tables

2.1	A Survey on Citation Reason Classification after Year 2005	19
2.1	A Survey on Citation Reason Classification after Year 2005 (Continued)	20
2.1	A Survey on Citation Reason Classification after Year 2005 (Continued)	21
2.2	XIP’s Annotations	25
3.1	Sentiment Polarity Distribution	35
4.1	Identified Key Terms for on All Possible Citation Reason Classes	45
4.1	Identified Key Terms for on All Possible Citation Reason Classes (Continued)	46
4.1	Identified Key Terms for on All Possible Citation Reason Classes (Continued)	47
4.1	Identified Key Terms for on All Possible Citation Reason Classes (Continued)	48
4.1	Identified Key Terms for on All Possible Citation Reason Classes (Continued)	49
4.2	Citation Reason Properties	51
4.3	CCRO Evaluation using Automated Tools	57
4.4	CCRO’s Evaluators Profiles	59
4.5	CCRO Evaluation by Domain Experts	60
5.1	Verbs and No of Times Used in All 3 Sentiments	67
5.1	Verbs and No of Times Used in All 3 Sentiments (Continued)	68
5.2	CCRO’s Annotators Profiles	76
5.3	CCRO Properties Distribution	78
6.1	SPARQL Query 1: Results	91
6.2	SPARQL Query 2: Results	92
6.3	SPARQL Query 3: Results	94
6.4	SPARQL Query 4: Results	95
6.5	SPARQL Query 5: Results	97
7.1	Summary Statistics of use of L ^A T _E X in Science Disciplines	99
7.2	Summary Statistics of use of L ^A T _E X in Science Disciplines	100
1	Identified Citation Reason	129

1	Identified Citation Reason (Continued)	130
1	Identified Citation Reason (Continued)	131
1	Identified Citation Reason (Continued)	132
2	Reporting Verbs	134
3	Questionnaire for User-based Evaluation of CCRO	143
4	BiRO, the Bibliographic Reference Ontology	147
5	C4O, the Citation Counting and Context Characterization Ontology	148
6	FaBiO, the FRBR-aligned Bibliographic Ontology	149
7	DoCO, the Document Components Ontology	149
8	PRO, the Publishing Roles Ontology	150
9	PSO, the Publishing Status Ontology	151
10	SWAN, Discourse Ontology	151
11	CiTO, Citation Typing Ontology	152

Abbreviations

CCRO	Citation's Context and Reasons Ontology
CiTO	Citation Typing Ontology
CiTO-Ps	CiTO Properties
MAKG	Microsoft Academic Knowledge Graph

Symbols

V_p^i i^{th} verb in “Positive” sentiment

V_n^i i^{th} verb in “Negative” sentiment

V_o^i i^{th} verb in “Neutral” sentiment

F_p^i Frequency Percentage of i^{th} verb in “Positive” sentiment

F_n^i Frequency Percentage of i^{th} verb in “Negative” sentiment

F_o^i Frequency Percentage of i^{th} verb in “Neutral” sentiment

V_R Reporting Verb

Chapter 1

Introduction

1.1 Citation and Citation Graph

A reference to a published source or even an unpublished one is known as a citation. Citations create relationships between studies and drive new researches [1]. Authors of these scientific researches use citations as a fundamental element [2] to critically analyze other researches and to support their own. Therefore, citations become a vital part to establish relationships between publications [3]. According to Small [4], citing is a process of creating cognitive links between concepts, procedures, types of data, and documents. This view also echoes Garfield's [5] notion of cited documents. As Cronin [6] states, "citations are frozen footprints in the landscape of scholarly achievement; footprints which bear witness to the passage of ideas" (p. 16).

The network of Scientific Literature contains a collection of information entities (research papers) inter-connected by a link structure. "Citation" is the link between research papers to form an interconnected network, known as "Citation Graph". In this graph representation, each paper becomes the node and their citation link becomes the edge between the nodes; the edge implies that the paper associated to the first node is cited by the paper at the other node (past to present in the time domain), thus making a directed acyclic graph (DAG). The citation

graph may contain valuable information regarding how a scholarly activity has evolved during its lifecycle. Analysis and mining of the citation graph are used to extract this valuable information. Today's citation graph analysis mainly uses these citation links without any cognitive relationship between the nodes.

Citations allow authors and readers to make selections in several contexts at the same time. The interaction between authors and readers creates a sub-textual code of communication, thus inducing the perception of cognitive dimension in scientific literature [7]. The specific context or reason represented through citations from the perspective of the author defines the cognitive relationship between the citing and the cited paper. Enrichment of citation graphs with cognitive relationships using semantic tags can open new avenues for qualitative analysis in the field of scientometrics [8].

1.2 Citation Analysis

Various bibliometric methods have been used in the past to trace relationships between scientific research papers. Using these relationships, a knowledge structure can be mapped with the help of methods like "citation analysis", "bibliographic coupling", "co-word analysis" and "co-citation analysis". There are many applications based on these bibliometric studies [9, 10]. However, these studies are quantitative. They have their strengths but there are some limitations [11]. A few of these limitations are:

1. Unable to discover the nature of the relationship between the research papers;
2. Two research papers are considered relevant if one has cited the other irrespective of the research area that the papers belong to;
3. Sometimes citations are for citation purposes only known as "ceremonial citations" and can provide a certain research paper with undue popularity.

In our humble opinion, two types of researches exist in the field of citation analysis to evaluate research papers; one is defined using citations' count, to evaluate the impact [5], and the other type is defined by its content, known as "Citation Content Analysis" [12]. Citation Content Analysis (*CCA*) is a method to analyze citation meanings because it comes from the nature of academic writing itself. It has been accepted and validated that citation content analysis is most efficient when applied to semantically rich and logically consistent texts [13]. Academic writings such as research papers meet all these requirements since they are formal, official, systematic, and neutral to a great degree. Another reason, why *CCA* should be used to investigate citing behaviors, is embedded in the symbolic nature of citations itself.

Though citation analysis is multidisciplinary research where it can be studied from various aspects, from information science (bibliometrics) to linguistics (discourse analysis) and from separate orientations - quantitative and qualitative. Information scientists focus on the frequency of a citation, while linguists focus on the embedded meaning of a citation. However, we believe that information science can benefit from the discoveries of citation analysis in the linguistic domain. Therefore, this research deals with the qualitative analysis of citations using well-defined techniques in the area of semantic computing, ontology engineering, and knowledge graphs by taking into account the rhetorical and linguistic choices of the author.

1.3 Citation Context and Reasons

There is a difference between a citation and reference [14]. The reference refers to the works mentioned in the reference section or bibliography of a journal article. A reference may be mentioned once or multiple times in an article. Each mention is considered as a citation. Thus, citations are the contexts in which references are made. There are many studies that are based on the context or sentiment of a citation [15–21]. In general a citation can have "Positive", "Negative" or "Neutral"

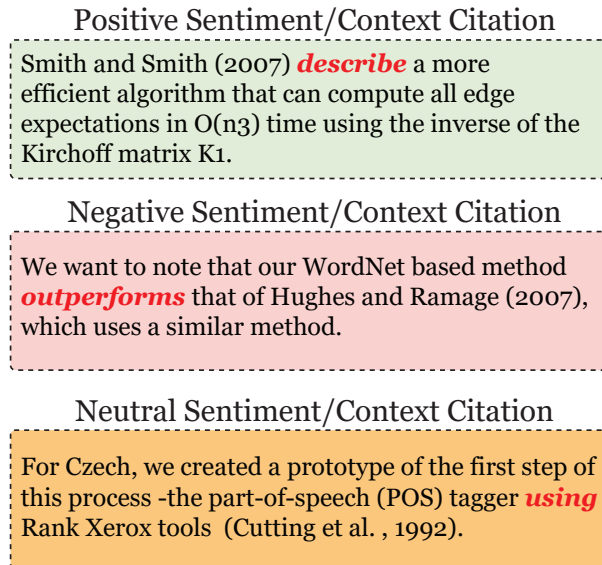


FIGURE 1.1: Citation's Context/Sentiment Examples

sentiments or contexts as shown in Fig 1.1. Though the citation sentiment analysis can provide a simple context for citations in which references are made. But there can be various reasons, why an author cites another research beyond simple sentiment analysis. A citation link in today's citation graph mainly implies that the paper at the other node cites the paper associated with the first node, without showing any cognitive relationship between the two. Thus, there is a possibility to enrich these citation graphs with cognitive relationships using meaningful and semantic tags for citation reasons.

A citation graph can be considered as a forest of graphs with millions of individual graphs and hundreds of nodes in each graph. According to a study to estimate the size of data need to be processed for citation graph in 2014 [22], there are over 45 Million research papers in Microsoft Academic Research, over 55 Million in Web of Science, and over 100 Million in Google Scholar. Annotating each citation in such a large volume of citation data in one of the citations' reasons is nearly impossible. Moreover, the citation's reasons have overlapped and diffused meanings. Discovering these citations' reasons using machine algorithms will also result in very low accuracy. However, a disjoint and formally defined set of citation's reasons can make machine algorithms and automated reasoners to identify these citation's reasons with high accuracy.

1.4 Open Citations

Using *OWL2* ontologies [23], it is now possible to encode bibliographic and citation data, document components, and the nature of individual citations into formal and machine-interpretable form. There have been many experiments and researches to encode various components of a scientific paper in a formal way including citations using ontologies. One of such efforts is *SPAR* (Semantic Publishing and Referencing) Ontologies [24]. *SPAR* contains variety of different ontologies. One of the most relevant ontology to our research has been developed by Peroni and Shotton [25], introduced as *CiTO*, an ontology to assert and characterize bibliographic reference and citations. *CiTO* describes 41 properties, termed as *T41*, to define and annotate a citation with its reasons. Annotating 41 properties also requires a huge cognitive effort [26], therefore, there is a need for a smaller set of properties to make it more effective. Moreover, a mechanism is required to integrate these citation reasons while authoring a research article.

To facilitate open citations, numerous projects are described in literature such as *DBLP in RDF* [27], *Springer SciGraph*¹, and *OpenCitations* [28]. Though these projects provide state-of-the-art semantic web technologies to model bibliographic data, they have their limitations as well. *DBLP in RDF* is restricted to only one discipline, *SciGraph* deals publications from a single publisher, and *OpenCitations* create citation relations without considering the entity type of a cited document. Recently, the emergence of Knowledge Graphs to represent a collection of inter-linked entities via semantic metadata has led researchers to represent open citations in a knowledge graph. A few worth mentioning initiatives are *WikiCite*², *AceKG* (A Large-scale Knowledge Graph) [29], and *MAKG* (Microsoft Academic Knowledge Graph) [30, 31]. The most prominent among these is (Microsoft Academic Knowledge Graph - *MAKG* with 209,792,741 papers, 1,380,196,397 references, and 8 Billion triples. However, *MAKG* uses *CiTO* Ontology to encode bibliographic data. As *CiTO* has its limitations (Section 2.2), these limitations become an inherent part of *MAKG* as well.

¹Springer SciGraph: <https://www.springernature.com/de/researchers/scigraph>

²WikiCite: <http://wikicite.org/>

1.5 Scientific Authoring of Citation Reasons

Authors cite other researches to make claims on their findings or to base their models on certain findings or simply contradict or negate results [32]. However, today's available research article authoring tools do not provide a mechanism for modeling and annotation of citation reasons. Microsoft Word and L^AT_EX are the most used authoring framework for writing research papers. L^AT_EX is a de facto standard in scientific communities such as Physics, Mathematics, and Computer Science for scientific authoring and publishing while other communities such as Biomedical prefer Microsoft Word instead [32]. The study suggests that available plugins for citation in Microsoft Word and the citation packages for L^AT_EX, both are focused on the formatting of citation in various styles and do not provide support for authors to integrate citation reasons while authoring the research document. Integration of citations' context and reasons using semantic web technologies within these authoring tools, can help authors to specify a reason while citing other articles.

Semantically Annotated L^AT_EX for Scientific Publications - SALT [33] is a semantic authoring framework that targets enrichment of scientific publications using three ontologies. The ontology that captures the rhetorical and argumentation structure of a research article is the Rhetorical Ontology. Using this ontology, SALT provides plugins for both Microsoft Word and L^AT_EX to manually annotate a research document using semantic tags. However, the SALT only provides a mechanism to model rhetorical and argumentative structure within a scientific document while enrichment of citations' reason requires modeling of a rhetorical structure across multiple scientific documents.

1.6 Motivation

Isaac Newton once said, "I can see further because I stand on the shoulders of giants." Authors of academic papers perch themselves on the shoulders of others to

develop an argument to support their claim, shaping an evolutionary path within a scholarly activity. Today's digital libraries have become the digital replicas of paper artifacts [32] where automatic analysis of such evolutionary paths becomes difficult due to the absence of citation reasons in the replica. According to Shum [32], digital libraries need to answer queries like;

1. What is the current state of the debate on this question?
2. Who disagrees with this theory?
3. Was this prediction ever fulfilled?
4. What assumptions does this approach depend on?
5. Are there different schools of thought around this problem?

However, due to the absence of citations' context and reasons, current digital libraries either provide little or no support towards these queries. Integration of citations' context and reasons using formal and meaningful tags, it is possible to discover, analyze, and infer evolutionary paths of scholarly activity against each citation reason.

1.7 Problem Statement

Due to a huge volume of available scientific literature in each domain, authors spend a lot of time to search and filter the required information. Availability of citations' context and reasons can help in minimizing the time and effort to find and filter research papers. Though an automatic extraction of these citation's context and reasons from existing literature becomes important, authoring a new research article without destroying the useful semantics can also help automatic reasoners to filter the required information. However, both the problems requires the development and adoption of a model (ontology) for citation reasons between research papers that permit bibliographic and citation's reason data in machine-readable *RDF* form. Therefore, a formal problem statement can be stated as

A minimal and disjoint set of citations' reasons between research articles is not available. There would be a need to semantically model these citations' reasons in a machine-understandable format (Ontology).

This model can be integrated with a futuristic research paper authoring tool, where an author can annotate the citations' reasons and can also be instantiated from the existing literature by using automatic reasoners. After a successful inclusion of these reasons, an application can be demonstrated with the evaluation of research evolutionary paths.

To solve the problem, at-least the following research questions are required to be answered.

1.8 Research Questions

- 1. What are the shortcomings in the existing semantic representations for the Citation's Context and Reasons?** (Chapter 2 - Literature Review)
- 2. What is the methodology to develop an ontology of Citation's Context and Reasons that are minimal and disjoint? What are its possible applications?** (Chapter 3 - Methodology)
- 3. What is the formal definition of Citation's Context and Reasons Ontology? What are the methods to evaluate and validate the proposed ontology?** (Chapter 4 - Ontology Development)
- 4. What is the procedure to instantiate and map the Citation's Context and Reasons Ontology on real data? What are its outcomes?** (Chapter 5 - Ontology Instantiation and Mapping)
- 5. What type of semantic queries can be applied on mapped data using Citation's Context and Reasons Ontology?** (Chapter 6 - Query Semantic Graph using CCRO)

6. **How can the developed ontology be integrated within the current research paper authoring tools?** (Chapter 7 - Semantic Authoring using CCRO)

Chapter 2

Literature Review

Automated processing of bibliographic and citations' data requires machine-interpretable metadata for publications and citations while the ontologies are required to encode these metadata elements [25]. Some of the areas that require answers are:

1. Development and adoption of semantic models (ontologies) that permits bibliographic and citations' reasons data in machine-interpretable form, is the core requirement in scholarly authoring and publishing;
2. Development of annotation tools to help the authors to enhance the semantic relation of their documents with others, using appropriate semantic assertions for the citations.

Semantic-based publishing applications provide customization of data and the content to reflect the user's needs of retrieval of relevant data with minimal effort. Using the new set of OWL2 ontologies [23], bibliographic and citation data, document components, and the nature of individual citations can be structured. However, existing applications do not follow the basic principle of semantic-based publishing as defined by Peroni and Shotton [25]. Our study reveals that such applications use metadata elements such as Authors and their affiliations, editors and their affiliation, publishing companies, etc., and do not look for the citation reasons. To understand semantic-based citation reasons, we have distributed our survey into

three parts. First, we have examined various available ontologies (Appendix 10) to find whether they provide options to record the citations reasons in a semantically meaningful way (minimal set of disjoint reasons) and what are their limitations (Section 2.1). Second, we have explored citation reasons to understand their types and classification in detail (Section 2.3). Kindly note, the studied literature contains both manual and automatic annotation of citations, however, the focus of this survey is to find and explore the granularity of citation reasons and not to look for machine-learning models for automatic annotation. Last but not least, we have investigated existing and available authoring tools and packages against the provision of integration of citation reasons while authoring a research article (Section 2.4).

2.1 Semantic Representations

2.1.1 BiRO - Bibliographic Reference Ontology

The Bibliographic Reference Ontology (*BiRo*) [34], based on *FRBR* [35] describes individual bibliographic reference and its relationship to the cited article using two properties; “is referenced by” and “reference” with domain and range as “endeavor” and “bibliographic record” alternatively. It is clear from the meanings that both properties neither define the nature of the relationship between the papers nor citation reasons.

2.1.2 C4O – Citation Counting and Context Characterization Ontology

The Citation Counting and Context Characterization Ontology (*C4O*) [34] keeps track of the number of citations that a paper has received using all possible external sources. The ontology claims to record the “context of citation”, however, this is an in-text reference pointer of where the citation has been made. Its “has

context” property provides the place where a possible rhetorical motivation for citation exists in a paper but does not exploit the context for possible motivation or reasons for citation.

2.1.3 FaBio – FRBR-aligned Bibliographic Ontology

The FRBR-aligned Bibliographic Ontology (*FaBio*) [25] mainly records publications such as books, magazines, journals, and their content like algorithms, specifications, vocabularies or technical reports, that are published or in the process of being published, using semantic web descriptions. It is based on *FRBR* [35] data model to interlink manifestations, items, and expression and does not deal with the nature of links between them (citations).

2.1.4 DoCO – Document Component Ontology

The Document Component Ontology (*DoCO*) [36] decomposes a research paper-document into its structural and rhetorical components such as Abstract, Introduction, Results, Conclusion, and Bibliography, etc., and stores these components using *RDF* - Resource Document Framework. The nature of a citation is the discourse element of a research paper and this ontology does not deal with it.

2.1.5 SWAN – Discourse Ontology

SWAN 1.0 Discourse Ontology [37] is designed to create an ecosystem that can create, store, access, integrate and exchange semantic context of scientific papers especially in the field of Neuro-medicine and specifically Alzheimer Disease (AD). The ontology stores a research statement with three possible discourse elements: “citeAsEvidence”, “citeLifeScienceEntity” and “citesReagent”. These discourse elements relate to each other using a set of relationships that are “discusses”, “refutes”, “supports” and “alternativeTo”. The ontology uses standard biological concepts [37] such as “genes”, “proteins”, “reagents” etc to assert scientific

discourse. Therefore, the ontology works fine in its intended domain but is not helpful in other domains. However, a smaller set of discourse elements provided by the ontology is helpful for the annotators.

2.1.6 PRO – Publishing Role Ontology

The Publishing Role Ontology¹ (*PRO*) stores the roles of agents such as people, organizations, or groups involved in the publication process. It also records the time when a role asserts. However, it does not deal with the citation or its nature.

2.1.7 PSO – Publishing Status Ontology

The Publishing Status Ontology² (*PSO*) records the status of research paper document during its life cycle. It also records the time duration the document took to transit from one status to another and the people involved during that. This ontology also does not deal with citations.

2.1.8 CiTO – Citation Typing Ontology

The Citation Typing Ontology (*CiTO*) [25] asserts and characterizes bibliographic references and citations. Citations have three characteristics “direct and explicit”, “indirect” and “implicit”. Based on biomedical researchers, the ontology describes citation nature in terms of the “Factual” and “Rhetorical” relationships and sub-divides them between “Positive”, “Negative” and “Neutral”. In total, there are 41 properties and are known as *CiTO-Ps*. A study [26] has been conducted to cluster these properties that exhibit similar meanings according to the subject’s annotation using the Chinese Whispers clustering algorithm [38]. The results show that a certain collection of properties show diffused and overlapped meanings.

¹<https://sparontologies.github.io/pro/current/pro.html>

²<https://sparontologies.github.io/psso/current/psso.html>

By examining the above ontologies, it becomes clear that the ontology that comes closest to our research goals is *CiTO*. It defines the nature of citations for intelligent linking and reasoning. However, the characterizations defined by *CiTO* are very difficult for humans to understand and adopt. Ciancarini [26] has summarized some problems in it after a careful analysis of both experimental data and subjects' feedback. Based on these and other experiments, some of the limitations in *CiTO* are:

2.2 CiTO Limitations

2.2.1 Less Used Properties

There is a number of properties defined in *CiTO-Ps* that are never used. For instance, the properties defining negative reasons such as “disagreesWith”, “disputes”, “parodies”, “plagiarizes”, “refutes”, “repliesTo”, “ridicules”, etc. are used less frequently than neutral and positive ones [39].

2.2.2 Most Used Neutral Properties

Some defined properties share different scholarly domains such as “citesForInformation” and “citesAsRelated” and are the most commonly used properties being the most neutral ones. Ciancarini [26] revealed that these two properties are most commonly used even for those instances that can be defined in a more precise manner such as “citesAsAuthority”, “citesAsDataSource”, “discusses”, etc.

2.2.3 Lower Inter-Rater Agreement

There are 41 properties in *CiTO* to define and annotate a citation for its reasons. Annotating 41 properties requires a huge cognitive effort. An experiment [26] between *T41* and *T10* was conducted where *T41* uses all *CiTO-Ps* (*CiTO*

Properties) and *T10* uses only a subset of 10 *CiTO*-Ps. The experiment revealed that the smaller set of properties are more usable as compared to a full set for annotating citations between professors, academic researchers, postdoc, and Ph.D. students.

2.2.4 Non-Taxonomic Organization of CiTO-Ps

CiTO does not follow a taxonomic organization. Each property has its own mapping determined by the mental model. Some *CiTO*-Ps exhibit a similar model and can be clustered into a parent property.

2.2.5 Customized Properties

CiTO lacks support for customization. If an annotator does not find a property that perfectly fits, her need; the annotator selects the property closest to her mental model. The latest *CiTO* release [25] has targeted this issue by making the organization of the ontology (i.e., the *TBox*) static, while users are free to express their characterizations precisely, capturing details and tones.

2.2.6 Misinterpretation of Properties

There are some properties in *CiTO* that users normally misunderstand, or interpret them in different ways, making a clear indicator for the need of improvements [26].

2.2.7 Properties Perspective

CiTO properties conform to the annotator's perspective and not to the author's perspective. "disagreesWith", "disputes", "parodies", "plagiarizes", "refutes", "repliesTo", "ridicules", etc. are some examples that only an annotator can use

and not the author himself. An author of a scientific paper can better define a citation reason than the perception of an annotator. This change of perspective can make the citation reasons more semantically defined.

Based on biomedical researchers, *CiTO* describes the nature of citations in scientific research articles in terms of factual and rhetorical relationships. However, there are a number of problems, especially the large number of properties defined in *CiTO* and the fact that all classes have the same weightage. Experiments have revealed that a lesser number of properties have better cognitive value and a taxonomic hierarchy between these properties is required. Therefore, a requirement emerges for the development of an ontology that focuses on the author's perspective and has a lesser number of properties for better cognitive value. We have sketched a methodology to define this new semantic representation of citation reasons known as "Citations' Context and Reasons Ontology" (Chapter 3) and its formal definition both at schema and instance level (Chapter 4). However, to develop this ontology, we need to understand the citations' reasons and their classification in detail.

2.3 Citation Reasons Classification

As citation counts have matured towards a serious means to assess the impact of a scholarly work, their applications have given rise to the criticism about the quality of simple counts. Therefore, various attempts have been made to examine the citation content or context for a deeper insight into scientific knowledge. The creator of citation indexes [40] has emphasized that the use of citations to evaluate a paper was not wise. He also argued that citation frequency can only measure the extent of research activity and not the significance of an author's work. Therefore, other techniques must be used to measure performance evaluations by classifying citations. One of the first proposed classification [41] defines: *for a single complex citation function, four values could be in use: "Conceptual", "Perfunctory", "Evolutionary" and "Negational"*. Later on, Garfield [42] identified 15 citation reasons;

why an author chooses to cite another paper. Simone Teufel [43] initially defined her citation categories to be mutually exclusive. Afterward, with the help of her colleagues, she converted the categories to a “Positive”, “Negative”, and “Neutral” scheme by defining a relationship between the binary sentiment classification and the citation function classification [44]. Athar [45] provided simplified citation categories using sentiment analysis and reduced the categories into just three classes: “Positive”, “Negative”, and “Neutral”. This resulted in more robust results than the previous classifications that used more classes. Dong and Schäfer [46] defined a citation classification schema with four types: “Background”, “Fundamental Idea”, “Technical Basis”, and “Comparison”. Li and his co-authors [47] grouped their citation function into three general categories: “Positive”, “Negative” and “Neutral” as well. Valenzuela [48] has described the importance of citation mining in a relatively different way. According to the authors of the paper, citation in a publication has different roles in knowledge diffusion. Frequency and the placement of citations in a section of a paper can have importance. Moreover, why a citation is made, was coarsely divided into two classes named as “Incidental” and “Important”. Recently, Taşkın [20] has defined “Meaning” and “Purpose” of a citation differently. The “Meaning” defines a citation with Positive, Negative, or Neutral sentiments but why an author has highlighted a citation is the “Purpose”, identified by five classes. In general, there are two types of research studies, citation’s sentiment based, and citation’s reason-based. However, we believe that the sentiment, meaning, and the purpose or reason of a citation are interlinked. A “Purpose” can have sentiment value. Therefore, a combination of both (Sentiment and Reason) in a taxonomic hierarchy needs to be adapted where each citation can initially be classified with a sentiment value and then by citation reason within the selected sentiment. Using this concept, CCRO provides a minimal set of citation contexts and reasons that are disjoint with a formal and meaningful definition of these reasons.

We have investigated various citation classifications and performed an analysis of the literature available after year 2005 only using the following benchmarks (Table 2.1).

1. **Properties:** What is the number of properties used to classify a citation?
2. **Type of Property:** What are properties based on? Sentiment or Cognitive
3. **Disjoint:** Are the properties used to classify citations disjoint in nature?
4. **Ontology:** Is there any formal definition available for citation reasons?
5. **Corpus:** What type of corpus is used for citation classification?

The survey reveals that *40%* of researches have used sentiment-based properties and *60%* have used cognitive-based properties for citation reasons. Sentiment-based properties (Positive, Negative, and Neutral) are naturally disjoint. However, cognitive-based properties may or may not be disjoint. Only three cognitive-based researches have citation reason properties that are partially disjointed while the rest of them have citation reason properties with diffused and overlapped meanings. A possible disjoint and formally defined set of reasons can make machine-learning algorithms to identify these with high accuracy. Another aspect that is evident from the survey is, no experiment has used citation reasons that are defined in a formal way in the form of an ontology. Thus, a disjoint and formally defined semantic citations reasons (if possible) that can integrate both sentiment and cognitive nature of citations can play a vital role.

Analysis, extraction, and classification of citations' reasons in scholarly documents have an important role in finding relationships between them. However, a methodology needs to be adapted where future research is documented in a more meaningful way, rather than a simple link between two documents. Let's investigate some of the existing and available authoring tools and packages against the provision of integration of citation reasons at the time of authoring a research article.

TABLE 2.1: A Survey on Citation Reason Classification after Year 2005

Sr.	Author	Properties	Sentiment	Cognitive	Disjoint	Ontology	Corpus
1.	Teufel et al. [39]	7	×	✓	×	×	CmpLg: Computation and Language
2.	Teufel et al. [43]	4	✓	×	✓	×	Computation and Language E-Print
3.	Angrosh [49]	4	✓	×	✓	×	Lecture Notes in CS (LNCS)
4.	Athar [45]	3	✓	×	✓	×	AAN - ACL Anthology Network
5.	Dong & Schäfer [46]	3	×	✓	×	×	AAN - ACL Anthology Network
6.	Athar & Teufel [44]	3	✓	×	✓	×	AAN - ACL Anthology Network
7.	Tandon & Jain [50]	5	×	✓	×	×	Own Defined Corpus
8.	Jochim & Schütze [51]	3	✓	×	✓	×	AAN - ACL Anthology Network
9.	Li et al. [47]	12	×	✓	×	×	Own Defined Corpus
10.	Han, Eric & Martin [52]	3	×	✓	×	×	AAN - ACL Anthology Network

TABLE 2.1: A Survey on Citation Reason Classification after Year 2005 (Continued)

Sr.	Author	Classes	Sentiment	Cognitive	Disjoint	Ontology	Corpus
11.	Wan & Liu [53]	2	×	✓	×	×	AAN - ACL Anthology Network
12.	Hernández & Gómez [54]	7	×	✓	✓	×	AAN - ACL Anthology Network
13.	Xu et al. [15]	3	✓	×	✓	×	PubMed Central (PMC)
14.	Valenzuela [48]	4	×	✓	×	×	Own Defined Corpus
15.	Butt et al. [16]	2	✓	×	✓	×	Own Defined Corpus
16.	Kim & Thoma [17]	3	✓	×	✓	×	MEDLINE
17.	Kazi & Patwardhan [18]	3	✓	×	✓	×	ACM Digital Library
18.	Taskin [20] - Purpose	7	✓	×	✓	×	Turkish LIS Publications
19.	Taskin [20] - Meaning	3	×	✓	×	×	Turkish LIS Publications
20.	Alvarez et al. [19]	8	×	✓	×	×	AAN - ACL Anthology Network

TABLE 2.1: A Survey on Citation Reason Classification after Year 2005 (Continued)

Sr.	Author	Classes	Sentiment	Cognitive	Disjoint	Ontology	Corpus
21.	Jurgens et al. [55]	7	×	✓	×	×	AAN - ACL Anthology Network
22.	Khadidja et al. [56]	6	×	✓	×	×	AAN - ACL Anthology Network
23.	Meng [57]	8	×	✓	×	×	PubMed
24.	Cohan et al. [58]	3	×	✓	×	×	SciCite
25.	Qayyum & Afzal [59]	2	×	✓	✓	×	AAN & Own Defined Corpus
26.	Perier-Camby et al. [60]	7	×	✓	×	×	AAN - ACL Anthology Network
27.	Halil et al. [21]	3	✓	×	✓	×	Own Defined Corpus
28.	Zhao et al. [61]	6	×	✓	×	×	ACL / ARC / NIPS
29.	Suppawong et al. [62]	4	×	✓	×	×	Own Defined Corpus
30.	Mingyang et al. [63]	3	×	✓	×	×	Dataset: [48] / Semantic Scholar

2.4 Semantic Authoring

Scientometrics provides insight into scholarly documents and patterns within publications [64]. However, it provides a little support for the qualitative nature of scholarly writing. Authors cite other researches to make claims on their findings or to base their models on certain findings or simply contradict or negate results, commonly known as discourse analysis [32]. We have performed a survey towards the semantic authoring of scientific literature and is comprised of two parts. First, we have examined available semantic tools that provide mechanism form modeling and annotation of semantic authoring. Secondly, we have investigated different packages available for L^AT_EX that provide support for authors to integrate scientific discourse while authoring the research document.

2.5 Authoring Tools for Citation Reasons

2.5.1 Human Authoring and Annotation

One of the first applications to develop semantic hypertext for scholarly discourse was developed in 2001-04 and is known as “ClaiMaker³”. “Claimaker” is the part of “ScholOnto” Project [65, 66] that provides a research prototype for usability testing, modeling and system development issues [67]. Based on “Claimaker” model, “ClaiMapper⁴” was developed. It is a visual based hypermedia tool that can store a claim in research paper in the form of semantic triple. These triples can later be interconnected to form a chain of complex nodes and structures. Similarly, in 2008 another tool known as “Cohere⁵” was released. It is highly interactive and open source web interface using *RESTful APIs*. It provides facilities to tag semantic annotations such as problem, hypothesis, assumptions etc. using *RDF*. In 2015, *Research Articles in Simplified HTML - RASH Framework* [68]

³Claimaker - <http://claimaker.open.ac.uk/>

⁴ClaiMapper - <http://compendium.open.ac.uk/institute>

⁵Cohere - <http://cohere.open.ac.uk>

integrated *RDF* with *HTML* to provide set of specifications and a tool for academic authoring. *RASH* is a markup language that provides a restricted *HTML* with only 25 elements and a facility for validation, visualization, conversion, and enhancement. The 'conversion' feature uses *XSLT* to convert a *RASH* document to \LaTeX using *ACM ICPS* and *Springer LNCS* styles whereas 'evaluation' feature uses Document Component Ontology (*DoCO*) for automatic annotation of markup elements to structural semantics. *RASH* provides easy to use mechanism for semantic annotations however it uses only `cito:cites` [69] for citation and does not provide any markup for scientific discourse.

All the above-mentioned tools except *RASH* provide semantic annotation for scientific discourse independent of the authoring environment. However, the semantic authoring process suggests enriching scientific publications with explicit linear, rhetorical, and argumentation structures while authoring a research publication [32]. In this context, one semantic authoring mechanism was proposed by *SALT* [33]. *SALT*, also known as Semantically Annotated \LaTeX , provides plugins for both \LaTeX and MS Word, where the author can manually annotate semantic tags while authoring the research document.

\LaTeX enables the authoring of documents using a high-quality typesetting system. It also provides a series of commands in a programmatic manner to produce the formatting and styling for the text. Due to its familiarity with authors to write content for publication and its capability for semantic authoring, \LaTeX will be our focus of research instead of MS Word. Therefore, a deeper insight into *SALT*'s \LaTeX plugin reveals that it provides three different types of annotations.

1. **SALT Rhetorical Ontology:** The author can correspond to a chunk of text as a rhetorical block using this ontology. Two tags are used to define the rhetorical block that is "`\begin{motivation}`" and "`\end{motivation}`".
2. **Elementary Discourse:** These items refer to a smaller sized text chunk along with rhetorical relations. It used \LaTeX commands such as "`claim[ID]{ ... }`" and "`cause{CLAIM_ID:SUPPORT_ID}`".

3. **Argumentation Elements:** These elements also provide elementary discourse using the positioning of claims within a document. It uses “`position[ID][CLAIM_ID]{ ... }`” command for the said purpose.

All these discourse commands require the presence of identification elements such as `ID` or `CLAIM_ID` to create the rhetorical relations. The author, while writing the document using \LaTeX , must create, manage, and track these IDs. Therefore, it might become difficult for authors because \LaTeX doesn't provide such a provision.

Another application that allows semantic annotation of discourse relationships in the form of a hypothesis, claims, and evidence in the biomedical domain is known as the *SWAN Workbench* [37]. The application uses *RDF* Triples to model and store relationships. However, this application was later replaced by *SWAN Annotation Framework* that integrated text mining algorithms to override manual annotation.

2.5.2 Automatic Authoring and Annotation

Automatic annotation schemes normally use argumentative zoning to detect rhetorical blocks based on the author's language. One of the applications based on this principle is Xerox Incremental Parser (*XIP*) [70] to perform rhetorical analysis on scientific papers. *XIP*'s annotation for rhetorical blocks is shown in Table 6.1. *XIP* labels the sentences with annotation tags more rigorously as compared to the reader of the document. However, the annotation list requires more rhetorical functions to describe research problems.

Another framework that provides automatic annotation of scientific discourse is the *SWAN Annotation Framework (AF)*. The framework works in conjunction with the *NIH-supported Neuroscience Information Framework (NIF)* [37]. *AF* is a three-tier application with the client-tier provides embedded web interface, the middle-tier provides text mining functionality and the data-tier provides persistence using *Annotation Ontology (AO)* [71]. Kindly note, the *SWAN Annotation*

TABLE 2.2: XIP’s Annotations

Annotation	Description
SUMMARIZING	summarizing aims, claims, results, conclusions
BACKGROUND KNOWLEDGE	descriptions of previous ideas
CONTRASTING IDEAS	descriptions of ideas as contrasting
NOVELTY	descriptions of new ideas
SIGNIFICANCE	descriptions of ideas as being significant
SURPRISE	descriptions of ideas as being surprising
OPEN QUESTION	descriptions of open questions
GENERALIZING	descriptions of research trends

Framework does not use *SWAN* ontology and rather uses *Annotation Ontology* to make it orthogonal to any domain.

The Above survey reveals that only a handful of applications or researches are available that provide annotation for the scientific discourse nature of research articles. Furthermore, whether the application provides human annotation or automatic, the application only deals with the elements by finding claims and ideas within a single document. Scholarly activities evolve with a passage of time and there is a need to embed the inter-connected nature of scientific literature (citation reasons) at the time of authoring a research document. \LaTeX provides “`\cite{paperID}`” command to cite another research and is widely used in all types of \LaTeX based authoring tools. However, “`\cite{paperID}`” command only creates a hyperlink without any cognitive link between the citing and the cited paper. A semantic annotation integrated within a “`\cite{paperID}`” command can empower the author to integrate the context and reason to cite. There are some variations available for \LaTeX “`\cite{paperID}`” command as well. Let’s

investigate some of its available variations to find if they provide any provision of semantic annotation for citation reasons or not.

2.6 L^AT_EX Cite Packages for Citations

2.6.1 Cite Package

The Cite Package is the most basic package for citation in L^AT_EX started in 1998. It is mostly intended for well-formed numeric citations [72]. The package only needs one command “\cite{paperID}” and is the natural behavior of L^AT_EX. However, there is hardly any documentation available for the complete package. Even Sebastian Rahtz, a long-term contributor to L^AT_EX typesetting when trying to provide support for it in “hyperref”, had to give up trying to understand it [73]. But there are several packages developed based on the Cite Package.

2.6.2 Harvard Package

The Harvard Package [74] qualifies citations by using the grammatical function of the label in the sentence and provides several commands. For example; when a citation is a noun, it uses “\citenoun{paperID}”, and when something must be affixed, it used “\citeaffixed{paperID}” command. The package also provides “\citeyear{paperID}”, “\citename{paperID}” and “\possessivecite{paperID}” commands as well. However, no semantic-based command is available for citation reasons.

2.6.3 Achicago Package

The Achicago Package [75], aimed at the Chicago Manual of Style, provides several bibliographic elements but doesn't use typeset quotations such as “\small{ }” or “\emp{ }”. It also provides multiple command such as “\citeNP{paperID}”,

“`\citeYear{paperID}`”, “`\citeN{paperID}`”, “`\citeA{paperID}`” etc. However, the main emphasis of the package is on what it is that the citation needs and not why the citation is made?

2.6.4 Natbib Package

The Natbib Package [76, 77] is the definitive word on author-year bibliography styles with \LaTeX . Build upon the Harvard Package, it provides a set of customization possibilities. It provides various commands such as “`\citep{paperID}`”, “`\citett{paperID}`”, “`\citeyearpar{paperID}`” and “`\citeauthor{paperID}`” but does not provide plain “`\cite{paperID}`”. Most of the \LaTeX templates available on the internet, use the Natbib Package as a chosen family of styles for citation. However, the package just provides citation styling in a variety of formats and does not integrate the semantic nature of citation reasons.

2.6.5 Apacite Package

The Apacite Package [78] provides citations and references according to American Psychological Association rules. The package can be customized in several ways. The apacite citation commands include “`\cite{paperID}`”, “`\citeA{paperID}`”, “`\citeAuthor{paperID}`”, “`\citeYear{paperID}`”, “`\citeNP{paperID}`” and “`\nocite{paperID}`” etc. Using the Natbib package, it also provides support for Full and short author lists, masked citations, and ad-hoc citations. However, this package also does not provide any support for the semantic nature of citation reasons.

The analysis of the survey reveals that citation packages create multiple forms of citation styles by providing variations in the basic “`\cite{paperID}`” command. Survey also reveals that the most commonly used package in \LaTeX templates is the Natbib Package. However, no package has integrated semantic or meaningful tags to define the context or reason of citation. Authors cite another research

based on some context or reason. Therefore, it is evident to develop a citation package that can integrate functionality to empower the author, to add citation's context and reasons while citing other documents to create a semantic or cognitive link between the citing and the cited paper, and thus making it possible to study the evolutionary paths in the scientific literature.

2.7 Conclusion

In our quest to understand semantic-based citation reasons, we have investigated various ontologies for semantic representation of citation reasons. We have also explored various citations reason-classification schemes. The literature survey reveals that there are several citation classification techniques. Two main categories of these classifications are either polarity based with a maximum of three classes or a high-granularity category scheme. Fine-grained and high-grained categories are difficult to be used with automatic reasoners as the difference between the classes is either too subtle or high [19]. Therefore, in the past few years, automatic citation classification studies are focused on a medium-grained approach. As citation classification with medium-grained 7 to 10 classes has become popular, these approaches have defined these classes based on their mental model. Therefore, there is a need to collect and systematically cluster available citation reasons. A few studies [19, 39, 48] show that the hierarchical clustering of these citation reasons can result in more robust results. As sentiment-based citations' reasons are disjoint, they do not represent why an author chooses to cite. Therefore, a combination of both sentiment and medium-grained cognitive-based citation reasons that are minimal and disjoint (if possible) is needed to be developed and semantically modeled in a machine-understandable format.

With the advent of Knowledge Graphs, the research began for storing scientific data in large scale *RDF* format. One such effort is the development of *Microsoft Academic Knowledge Graph (MAKG)* [30] with a huge volume of 8 billion triples

and its availability on *Linked Open Data Cloud*. However, it requires the adaptation of various ontologies to encode different parts of a research article. For references, *MAKG* has modeled the citation information using a separate ontology *CiTO*. Due to the coarse-grained 41 properties of *CiTO*, *MAKG* has only used one `cito:citation` as an entity type and leaving the rest. Though *MAKG* believes that citation context for each reference is valuable information for various tasks such as citation recommendation and citation-based paper summarization [30]. Therefore, the development of a minimal set of cognitive-based citations' contexts and reasons in the form of an ontology becomes inevitable.

Current research paper authoring and annotation tools do not provide integration of semantic and meaningful tags to define the context or reason of citation. These tools and packages are either focused on finding and tagging claims within a single document or to provide formatting styles for citations. Therefore, a system is required where the semantic model (ontology) of citations' reasons can be integrated within a futuristic research paper authoring tool that can help an author choose a semantic and meaningful tag for citation.

Chapter 3

Methodology

Based on the observations and findings of our literature review, we derive a set of requirements for an ontology to define the citation's reasons. The coarse-grained requirements to develop the ontology are:

1. **Minimal Reasons** - Defining and annotating a citation on its reason require cognitive effort. Thus, a minimum set of citation reasons is required.
2. **Disjoint Reasons** - An overlapping and diffused citation reason will result in a low accuracy for automatic reasoners. Thus, the defined reasons are required to be disjoint (if possible).
3. **Citation Context** - Citations' reasons can have different weights based on the context in which the citation is made. Thus, different contexts in which citations are made can have a significant role in the effectiveness of citation reason classification.
4. **Citation Mapping** - To recognize and tag citations on their reasons, it is necessary to link textual relations to ontological properties. Therefore, the ontology needs to define a mapping between the citation text and ontological properties.

5. **Re-usability** - Any defined ontology requires interpretation of the represented concepts and is the basis for its re-usability. Thus, defined ontology needs to provide annotations for its cognitive conceptualization.
6. **Simplicity** - Since the purpose of developing any ontology, is to be widely adapted and reused, therefore, the defined cognitive conceptualization has to be as simple as possible.

To the best of our knowledge, there is no existing ontology for citation reasons that fulfills the above requirements. Based on the ontology development methodology known as Methontology [79, 80], we present the methodology for Citation's Context and Reasons Ontology using Stanford linguistic Beth Levin's [81] inventory of English verbs and knowledge base.

Fig 3.1 displays the main process blocks of our adopted methodology. It mainly consists of three blocks. The first block defines steps to develop the ontology itself whereas the second block outlines the steps to instantiate and map the ontology classes and properties on the selected corpus. The last block defines possible applications using the developed ontology. Details of individual processes are next.

3.1 Ontology Development

The First block of our methodology focuses on the development of the ontology. It mainly consists of three parts. The First part performs a survey to collect the primary data, required for ontology development whereas the second part performs qualitative data analysis on the gathered data to formulate ontology. The last part provides ontology evaluation and validation.

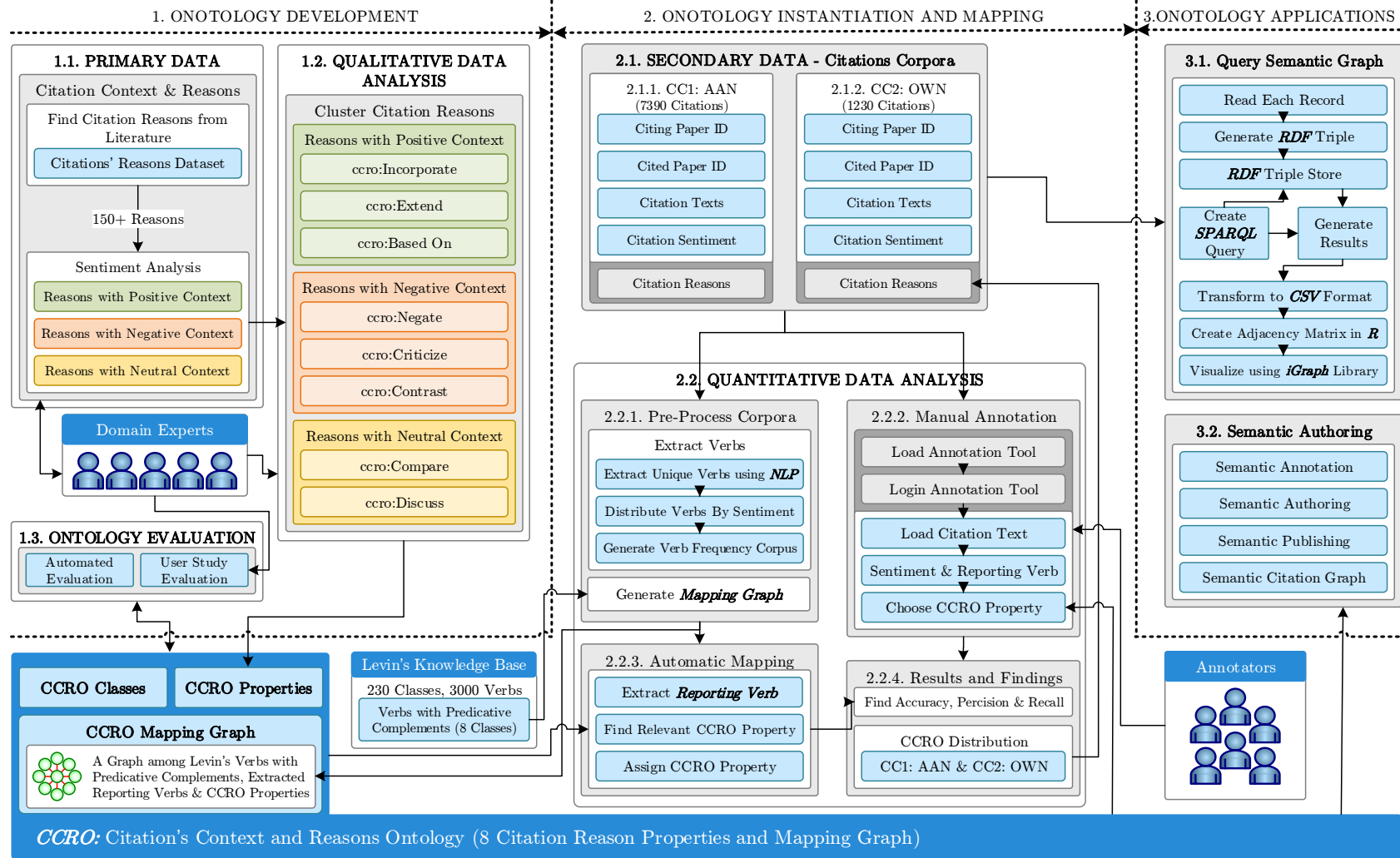


FIGURE 3.1: Methodology for Ontology Development and Mapping.

3.1.1 Primary Data - Citation's Context and Reasons

Primary data consists of terms found in published literature that define citations' reasons. There have been numerous researches for 5 decades to find the impact of a citation [82] or to recognize the type of influence [83] citations have. The first step in our methodology is to explore and list all such reasons and motivations found in the literature and classify them based on their sentimental context in "Positive", "Negative" and "Neutral" classes using an iterative process of sentiment analysis and experts' opinions. The resultant set of citation reasons (150+ Reasons - Appendix: A) will act as our primary data.

3.1.2 Qualitative Data Analysis - Cluster Citation Reasons with Titles

For a better understanding of primary data collected (150+ Citation Reasons), qualitative data analysis is performed to cluster citation reasons with similar meanings and motivations. As the collected citation reasons are large in number and have overlapped and diffused meanings. This step clusters the citation reasons with similar and disjoint meanings (if possible) using experts' opinions and qualitative data analysis techniques, and to name these newly formed clusters with a title exhibiting their collaborative meaning and constituting motivations. This set of classes will formulate the basis of our proposed ontology and will be stored as properties in the ontology.

3.1.3 Ontology Evaluation

Normally an ontology is evaluated and validated using two methods. One is using automated evaluation tools and other is by user study. Ontology evaluation process comprises of both methods. In automated evaluation, the ontology is evaluated against Syntax, Object and Data Properties, and Inconsistency Errors using world known automated tools. Whereas in user study evaluation, ontology

is evaluated against Incomplete Concept Classification, Disjoint Knowledge Omission, Exhaustive Knowledge Omission and Sufficient Knowledge Omission with the help of domain experts. The resultant is the ontology that is concise, complete and consistent.

3.2 Ontology Instantiation and Mapping

The next block of our adopted methodology focuses on the instantiation and mapping of developed ontology on real data. This process also consists of two parts. The first part focuses on the development of citations corpora whereas the second part performs quantitative data analysis on citation corpora.

3.2.1 Secondary Data - Citations Corpora

To instantiate and map citation's reasons using ontology, two citation corpora (*CC1* and *CC2*) are employed. Citation corpus *CC1* is publicly available and widely used in various similar experiments while *CC2* is our own formulated citation corpus. The two corpora are selected from two distinct areas of research. Citation corpus *CC1* is an archive of research articles in natural language processing and computational linguistics and citation corpus *CC2* is a collection of articles in computer science and scientometrics. Details of both corpora are delineated below.

3.2.1.1 Citation Corpus CC1:

To perform experiments, a standard corpus needs to be employed. The term "standard" means the selected corpus has been employed in similar experiments. Our survey suggests that the most common corpus is *ACL Anthology Network (AAN)* [84]. It is a comprehensive, manually curated corpus developed using papers published by ACL and Computational Linguistics journal for four decades.

TABLE 3.1: Sentiment Polarity Distribution

Sentiment Polarity	Corpus CC1	Corpus CC2
Positive	9.3%	28.2%
Negative	3.2%	12.2%
Neutral	87.5%	59.6%

In 2014, Athar [85] annotated the data of more than 8,700 citation sentences with “Citing Paper ID”, “Cited Paper ID”, “Citation Text” and “Sentiment Polarity”. This new annotated corpus contains “Positive”, “Negative” and “Neutral” citation texts and is publicly available. The corpus contains 9.3% citation texts with “Positive”, 3.2% with “Negative” and 87.5% with “Neutral” sentiments. We have used a subset of *AAN corpus* (7,390 Citation Sentences) to provide a complete mapping using the developed ontology’s constituent citation reason classes with a high inter-annotator agreement. Thus, adding a new column “Citation Reason” to *AAN Corpus*.

3.2.1.2 Citation Corpus CC2:

This corpus is a selected set of 40 random research papers, downloaded in *PDF* format from various sources. The selected articles come from four domains, with 10 highly representative papers from each domain. These domains are “H-Index”, “Scientometrics”, “Ontology” and “Sentiment Analysis”. Each paper is assigned a “Citing Paper ID” and their references are assigned “Cited Paper ID”. Annotation is citations based, therefore, citation sentences are manually extracted and curated as “Citation Text”. A Python-based script finds the sentiment of each “Citation Text” and places under “Sentiment Polarity”. The resultant corpus *CC2* becomes in line with *ACL Anthology Network corpus CC1*. *CC2* corpus contains 28.2% citation texts with “Positive”, 12.2% with “Negative” and 59.6% with “Neutral” sentiments. There is a difference in sentiment distribution in corpus *CC2* as compared to *CC1*. The corpus *CC2* is a careful selection of research articles

keeping in mind different schools of thought in the selected domain as compared to *CC1* where all the articles published in a journal are selected.

Available Online -

<https://github.com/imranihsan/CCR0/blob/master/CCR0-CC2.csv>

3.2.2 Quantitative Data Analysis

Both citation corpora have a large number of citation texts. To examine these, qualitative data analysis techniques are used that include a combination of automatic and manual annotation of data.

3.2.2.1 Pre-Process Corpora

Reporting verbs are the most important grammatical devices among all others, required to express a stance in an academic paper [86] and are used by authors to both report their claim and others [87]. The lexical and syntactic decision for the choice of reporting verb depends on the rhetorical context. According to Hopper [88], grammatical expression depends on the nature of the rhetorical context, and the choice of a reporting verb can not be made without understanding the rhetorical intent in which it is being used. Kindly note, the “*Reporting Verbs*” phrases used in this research, are the words used to report a claim by the author in a specific rhetorical context. In general, the sentence in which a reporting verb appears is a “Citation Sentence”. Several studies [44, 45, 51, 89] show that to find the rhetorical context of the citation sentence, sentiment analysis techniques can be used that classifies a citation in three possible classes: “Positive”, “Negative” and “Neutral”.

Part-of-Speech tagging is now possible using Natural Language Processing (NLP) Techniques. NLP techniques identify verbs and reporting verbs within a citation text. Combining the sentiment polarity of a citation and the reporting verb used in the context can provide a basis for complex problems such as finding the cognitive

relationship between a citing and cited paper or semantic enrichment of a citation graph.

Extract Verbs

Several tool-kits are available that can successfully recognize and tag part-of-speech verb clauses and link them with the knowledge base [90]. The survey suggests that the most common tool-kits are Google’s SyntaxNet [91], the Stanford CoreNLP Suite [92], the NLTK Python Library¹, and spaCy². We have selected spaCy for the experiment as it best suits our requirements and is substantially faster than other libraries. The process to tag and extract verbs from a citation text has three steps. The first step uses the POS Tagger to tag part-of-speech “Verbs” in the complete corpus. The next step uses Lemmatization and Stemming alternatively to convert verbs into its basic form. In the last step, unique verbs and their frequencies in the complete corpus is calculated. As both corpora contain citation texts available in all 3 sentiment polarities, therefore, all the extracted verbs are analyzed against the sentiment polarity of their parent citation texts as well. Thus, providing a corpus of verbs and their frequency distribution in each sentiment. A study suggests that different authors mostly use similar reporting verbs [93]. Therefore, the frequency of a verb can play a vital role in identifying a reporting verb. Verbs with high frequencies normally mean that they have been commonly used to cite a research article by different authors. Therefore, if a citation text contains more than one verb then the verb with the highest frequency will act as the Reporting Verb (Appendix B).

Generate Mapping Graph

A Stanford linguist Beth Levin [81] has provided an inventory of English verbs and organized it in a knowledge base. She has defined two distinct verb categories in the English language that are “transformation and creation” and “change of the

¹NLTK: <https://www.nltk.org/>

²spaCy: <https://spacy.io/>

state” that subsume several verbs. The complete inventory contains 230 classes and over 3000 English verbs, distributed among the defined classes. The class, we are interested in, is “Verbs with Predicative Complements” and has eight sub-classes with a unique set of properties and a collaborative meaning and each class has its collection of verbs.

Comparing the verbs in Levin’s knowledge base and the extracted verbs from different citation texts in each sentiment, a *Mapping Graph* can be formulated that provides a high level of abstraction on reason classes. Based on the citation context, one such property can be attributed to multiple classes. Therefore, the combination becomes a graph rather than a tree where one individual verb can belong to multiple classes, making the class semantically coherent.

3.2.2.2 Manual Annotation of Citation Reasons

For manual annotation of citation texts by experts in both corpora on citations’ reasons classes defined in the developed ontology, a tool is required. The annotation tool is a collaborative application, easily adaptable that provides a simple mechanism to annotate citation texts in both corpora *CC1* and *CC2* to citation reason classes. Each annotator is provided with credentials to access the tool. After login, the tools display one citation at a time for annotation. Annotators can navigate between various assigned citations using “Next” and “Previous” buttons. The tool shows “Citing Paper ID”, “Cited Paper ID”, “Citation Sentiment”, and “Citation Text”. It also highlights *Reporting Verb* in a citation text. The annotator is also provided with a subset of citations’ reason classes corresponding to the sentiment polarity of the citation text to choose the best-suited citation reason and submits to move onto the next annotation.

3.2.2.3 Automatic Mapping of Citation Reasons

This module assigns a citation text with one or more possible citations’ reasons. Using NLP techniques, *Reporting Verb* is extracted from each citation. Kindly

note that by *Reporting Verb*, it means that if a citation text has more than one verbs then *Reporting Verb* defines the most suitable verb used to relate the source with the author's claim. Next step uses the extracted *Reporting Verb* to assign a *CCRO* Property with the help of formulated *Mapping Graph*.

3.2.2.4 Results and Findings

To analyze the results, both automatic and manual annotation of citations' reasons on the selected corpora *CC1* and *CC2* are compared to find the Accuracy, Precision, and Recall. After a comparison of both automatic and manual annotation, a distribution of Citation Reason Classes on both Corpora *CC1* and *CC2* is formulated. Thus, adding a new column "Citation Reason" against each citation text within each corpus. Based on the results, both corpora can be visualized as a networked information space inter-connected by a link-structure. In the graph representation, each citing and cited paper becomes the node and their citation reason class becomes the edge between the nodes. In order to achieve this, citing paper and cited paper *IDs* from both corpora *CC1* and *CC2* are used along with the citation reason and its sentiment polarity. The complete data is then stored in a *CSV* format {Citing Paper ID, Cited Paper ID, Citation Sentiment, Citation Reason}. The resultant is a directed weighted graph between the citing and the cited papers with citation reasons as edge weight. The graph is then visualized for analysis.

3.3 Ontology Applications

According to Khabsa & Giles Method [22], an estimate based empirical data analysis revealed number of documents available on digital libraries from 1700 to 2014. According to study, in 2014, Google Scholar had 99.8 Million, Web of Sciences had 56.9 Million and Microsoft Academic Search had 45.9 Million documents. The amount in today's world has grown exponentially. Analysis and classification of citations in these documents has an important role. However, a methodology

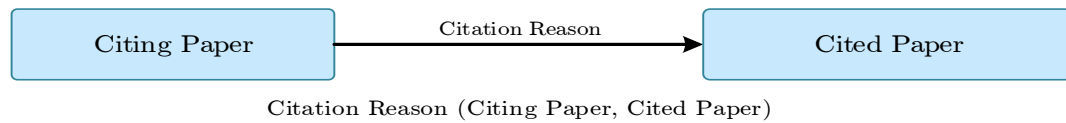


FIGURE 3.2: RDF Triple in RDF Data Store

needs to be adapted where future research is documented in a more meaningful way, rather than a simple link between two documents. Thus, there can be two (2) possible applications; one that can discover and query citation reasons from existing scholarly data and the other that can be used by authors to integrate these reasons at the time of authoring a research article. In our proposed methodology, both application are incorporated. First converts Corpus *CC2* into a Link Open Data (LOD) to formulate and query a semantic graph where as second defines the semantic authoring using features of \LaTeX based authoring.

3.3.1 Query Semantic Graph

After annotating both corpora on citation reason classes, an application is developed that reads each record, automatically extracts, and converts them into an *RDF* Triple. Resource Document Framework or *RDF* is a formal data model designed by *W3C*³ for machine-understandable metadata that stores the relationship between resources. The *RDF* structure uses *RDF* triples in the form of the subject (a resource identified by a URL), predicate (property), object (the value of the property). Using the *RDF* Triple composition “Subject – Predicate – Object”, each triple contains the Citing Paper as the “Subject”, the Cited Paper as the “Object” and the selected citation reason Class as the “Predicate” as shown in Fig 3.2. This collection of all *RDF* Triples is known as *RDF Triple Store*.

In the Semantic Web, *RDF Triple Store* is a directed and labeled graph. *SPARQL Protocol and RDF Query Language (SPARQL)* is an *RDF* query language that has capabilities for querying required and optional graph patterns along with their

³W3C RDF: <https://www.w3.org/RDF/>

conjunctions and disjunctions. Using *SPARQL* various semantic queries are formulated against different *CCRO*'s Properties for citation reasons. The resultant of a *SPARQL* query can also be an *RDF* Graph. These results are then tabulated and visualized as a directed labeled *RDF* Graph.

3.3.2 Semantic Authoring

One of the best sources of knowledge to tell the reason for a citation is the author of the paper at the time when he/she is writing the paper. The Semantic Authoring Application defines a Semantic Publishing Ecosystem that spans over four steps, “Semantic Annotation”, “Semantic Authoring”, “Semantic Publishing”, and “Semantic Graph”. A study [94] suggests that approximately 27% of researchers use \LaTeX based authoring tools. The numbers become astonishing when it comes to hard sciences such as Mathematics, where 97% of researchers use \LaTeX . Similarly, in Statistics and Probability, it is 89%, in Physics 74%, and Computer Science 46%. However, there is a community of researchers that believe in traditional processing systems such as Microsoft Word, Google Docs, LibreOffice, Apple Pages, etc. Due to the open-source \LaTeX document preparation system, we have selected \LaTeX based authoring to incorporate Semantic Publishing Ecosystem. \LaTeX also provides add-on features known as packages for adding new or modifying existing features programmatically. Therefore, for the first step in semantic publishing, “Semantic Annotation” defines “*CCRO*” Package with a semantic-based “`\cite`” command using “*Natbib*” Package for \LaTeX . “Semantic Authoring” integrates “*CCRO*” Package to generate Semantic \LaTeX files. “Semantic Publishing” reads semantic \LaTeX files and converts them into an *RDF* Triple Store whereas “Semantic Graph” provides query interface and visualization of the results.

3.4 Conclusion

The proposed methodology has three major parts. The first part defines the ontology using primary data (citation reasons) and qualitative data analysis with the help of domain experts. Chapter 4 defines experiments performed for primary data collection and formal definition of the proposed ontology using well-defined ontology development techniques [95]. The second part defines the mechanism to instantiate and map ontology on real data. In Chapter 5, the ontology is instantiated and mapped on real data with the help of case studies. The last part defines two different applications for the developed ontology. The first application focuses on providing a semantic query interface on existing literature (Chapter 6) whereas the second application defines a semantic authoring mechanism (Chapter 7) for future research.

Chapter 4

Ontology Development

Simple Knowledge Organization System (SKOS) [95] defines a framework for creating a formal ontology. The framework helps to identify key concepts in the domain of interest and their relationships. These concepts can be further enhanced using sufficient conditions using Description Logics (DL) [96] semantic model and Web Ontology Language (OWL). As the ontology evolves, there is a need for evaluation and validation of the ontology concerning the level of semantics incorporated in the ontology. Figure 4.1 defines steps for any ontology development using the methodology that enables the construction of ontologies at the knowledge level, known as Methontology [79, 80].

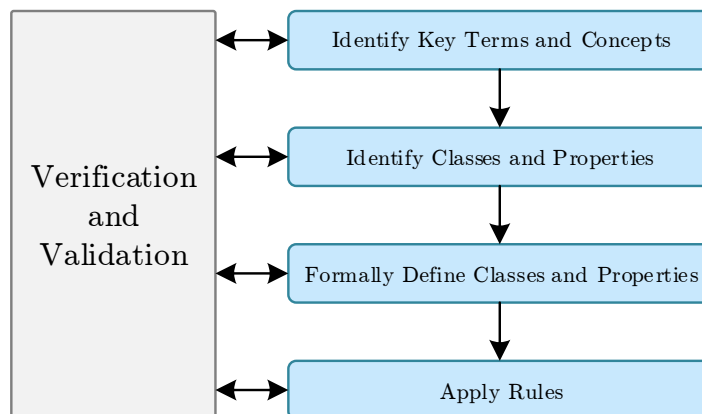


FIGURE 4.1: Ontology Development Process.

To develop the ontology, two different experiments are performed. First is knowledge acquisition by identifying key terms and concepts used for citation reasons available in the literature and classify them using sentiment analysis techniques. Second is ontology conceptualization by identifying and clustering classes and properties based on meanings and motivations within their respective sentiments with the help of experts. These experiments are performed by following the standard methodology for ontology development (Methontology) [79, 80], and their results are described in the coming sub-sections. The formal definition of identified classes and properties to formulate the ontology and its verification and validation are defined in Sections 4.3 and 4.4.

4.1 Knowledge Acquisition: Identify Key Terms and Concepts

Key terms and concepts, in the context of our domain, are terms and keywords extracted from the published literature that are used to define citation context and reasons.

4.1.1 Citation Context and Reasons

We have collected more than 150 citation reasons after studying the available literature (1965 – 2020) about the possible reasons to cite a research paper (Appendix A). This collection of citation reasons depicts that the reasons are not unique and several researchers have used similar keywords to explain their citation classes. So the citation reasons have overlapped and diffused meanings but different contexts (sentiments). Identifying key terms and concepts for knowledge acquisition is an iterative process. In this process, the first step is to broadly categorize terms in three distinct context classes; “Positive”, “Negative” and “Neutral” using sentiment analysis techniques. Identified key terms and concepts for citation reasons are outlined in a tabular form in Table 4.1.

TABLE 4.1: Identified Key Terms for on All Possible Citation Reason Classes

Sr.	Source	Reasons	Year	Positive Context Reasons	Negative Context Reasons	Neutral Context Reasons
1.	Liptez [97]	4	1965	Scientific Contribution Continuity Relationship Disposition of Contribution	-	Non-Scientific Contribution
2.	Chubin & Moitra [98]	6	1975	Affirmative Perfunctory Affirmative Subsidiary	Negative Partial Negative Total	Affirmative Basic Affirmative Additional
3.	Moravesik & Murugesan [41]	4	1975	Conceptual Evolutionary Perfunctory	Confirmative	-
4.	Spiegel-Rosing [99]	7	1977	Concept Point of Departure	-	Comparative History Data in Text Data in Tables Interpretation
5.	Frost [100]	5	1979	Factual Evidence Primary Text	-	View of Other Scholars Further Reading Previous Scholarship
6.	Oppenheim & Garfield [101]	7	1980	Theoretical Equation	Theory Not Applicable	Historical Background Data (Comparative) Data (Not Comparative) Relevant Work Methodology
7.	Pertiz [102]	6	1983	-	Argumentative	Setting Stage Background Comparative Documentary Methodology

TABLE 4.1: Identified Key Terms for on All Possible Citation Reason Classes (Continued)

Sr.	Source	Reasons	Year	Positive Context Reasons	Negative Context Reasons	Neutral Context Reasons
8.	McCain & Turner [103]	5	1989	Results Central Results Peripheral	-	Introduction-Central Introduction-Peripheral Methods-Central
9.	Eugene Garfield [42]	15	1996	Giving Homage Giving Credit Sustaining Claims Correct Own Work Correct Others Work Alert Forthcoming Work Authenticate Data Identify Idea Identify Concept	Criticize Disclaim Work Dispute Priority	Identify Methodology Background Lead to Uncited Work
10.	Teufel, Siddharthan, & Tidhar [39]	7	2006	Alternate Approach	Research Gap	Current Work Background Introduction Citation Sentences Descriptive Sentences
11.	Teufel, Siddharthan, & Tidhar [43]	4	2006	Positive	Contrast Weakness	Neutral
12.	Jörg [104]	6	2008	Inspired Extended Based Proposed	-	Compare Described
13.	Angrosh, Cranefield, & Stanger [49]	4	2010	Citation Positive Sentiment	Contrast Weakness	Neutral

TABLE 4.1: Identified Key Terms for on All Possible Citation Reason Classes (Continued)

Sr.	Source	Reasons	Year	Positive Context Reasons	Negative Context Reasons	Neutral Context Reasons
14.	Athar [45]	3	2011	Positive	Negative	Neutral
15.	Dong and Schäfer [46]	3	2011	Fundamental Idea Technical Basis	-	Background
16.	Athar and Teufel [44]	3	2012	Positive	Negative	Neutral
17.	Tandon and Jain [50]	5	2012	Strength Application	Limitations	Summary Related Work
18.	Jochim & Schütze [51]	3	2012	Positive	Negative	Neutral
19.	Yu [89]	5	2013	Neutral (Implicitly Positive) Positive	Negative Mitigated	Neutral
20.	Li, He, Meyers, & Grishman [47]	12	2013	Based on Corroboration Discover Positive Practical Significant Standard Supply	Negative Contrast	Co-citation Neutral
21.	Han Xu, Eric Martin [52]	3	2013	Functional Perfunctory	-	Hard to Tell

TABLE 4.1: Identified Key Terms for on All Possible Citation Reason Classes (Continued)

Sr.	Source	Reasons	Year	Positive Context Reasons	Negative Context Reasons	Neutral Context Reasons
22.	Hernández A. & Gómez [54]	7	2014	Based On Useful Correct	Debate Contrast Weakness	Acknowledge
23.	Wan and Liu [53]	2	2014	Strength Importance	-	-
24.	Valenzuela, Ha, & Etzioni [48]	4	2015	Using the Work Extending the Work	-	Comparison Related Work
25.	Butt [16]	2	2015	Positive	Negative	-
26.	Kim and Thoma [17]	3	2015	Positive	Negative	Neutral
27.	Xu [15]	3	2015	Positive	Negative	Neutral
28.	Kazi and Patwardan [18]	3	2016	Positive	Negative	Self-Citation
29.	Jha, Jbara, Qazvinian, and Radev [105]	6	2017	Use Sub-stain Basis	Criticize	Comparison Neutral
30.	Taşkın and Al [20]	5	2017	Method Data Validation	-	Literature Review Definition Data
31.	Alvarez et al. [19]	8	2017	Corroboration Based On Useful	Weekness Contrast Hedges	Acknowledgement Supply

TABLE 4.1: Identified Key Terms for on All Possible Citation Reason Classes (Continued)

Sr.	Source	Reasons	Year	Positive Context Reasons	Negative Context Reasons	Neutral Context Reasons
32.	Jurgens et al. [55]	7	2018	Motivation Uses Extension Continuation	Contrast	Background Future
33.	Khadidja et al. [56]	6	2018	Based On Useful	Contrast Weakness Hedges	Acknowledge
34.	Jia Meng [57]	8	2018	Confirmation Being Confirmed	Contrast/Conflict Unsolved	Background Statement Comparison Multi-Comparison Related Work
35.	Cohan et al. [58]	3	2019			Background Method Result Comparison
36.	Qayyum & Afzal [59]	2	2019	Important	-	Not Important
37.	Perier-Camby et al. [60]	7	2019	Motivation Uses Extension Continuation	Contrast	Background Future
38.	Halil et al. [21]	3	2019	Positive	Negative	Neutral
39.	Zhao et al. [61]	6	2019	Produce Use Extent		Introduce Compare Other
40.	Suppawong et al. [62]	4	2020	Use Extend	Notalgo	Mention
41.	Mingyang et al. [63]	3	2020	Utilize the Work Extend the Work		Comparison

Respondents to the first step in key term identification appear to be somewhat inefficient in terms of citation context. Not everyone has defined citations' reasons that fall in all three citation contexts (sentiments). Those, who have given reasons for all the three contexts, have not provided reasonable sub-classes. The next step towards ontology development is to cluster citation reasons in each sentiment context using their collaborative meanings with the help of experts' opinions. Based on our findings and experiments, we can formally define classes and properties for Ontology for Citation's Context and Reasons – *CCRO*.

4.2 Ontology Conceptualization: Identify Classes and Properties

4.2.1 Cluster Citation Reasons

After classifying the citation reasons in sentiment-based classes, the next experiment is to cluster the collection of citation reasons based on the meanings and constituting motivations. A group of experts that consists of four English Linguists (2 Lecturers and 2 MS Students) have created clusters of citation reasons within each sentiment. After the complete process, reconciliation is made to formulate eight clusters with an inter-annotator agreement of over 88%. The experiment has revealed there are three clusters of citation reasons in “Positive Reasons”, three in “Negative Reasons” and two in “Neutral Reasons”. Each cluster is provided with an appropriate title that exhibits the cluster's collaborative meanings. In semantic computing, knowledge graphs use semantic tags (properties) to create relationships between entities. Citation reasons are the relationships between citing and cited papers. Therefore, these eight clusters become properties. The new set of properties and their collaborative meanings are shown in Table 4.2, whereas formulated clusters are shown in Fig 4.2.

TABLE 4.2: Citation Reason Properties

Context	Property	Collaborative Meanings
Positive	Incorporate	To cite a research as part of a whole
	Extend	To spread from a central research to a wider solution
	Based On	To use a research as foundation or starting point
Negative	Negate	To cause to be ineffective or invalid
	Criticize	To find fault in a research with: points out the faults of
	Contrast	To show differences with opposite nature
Neutral	Compare	To examine in order to show similarities
	Discuss	To consider or examine by argument

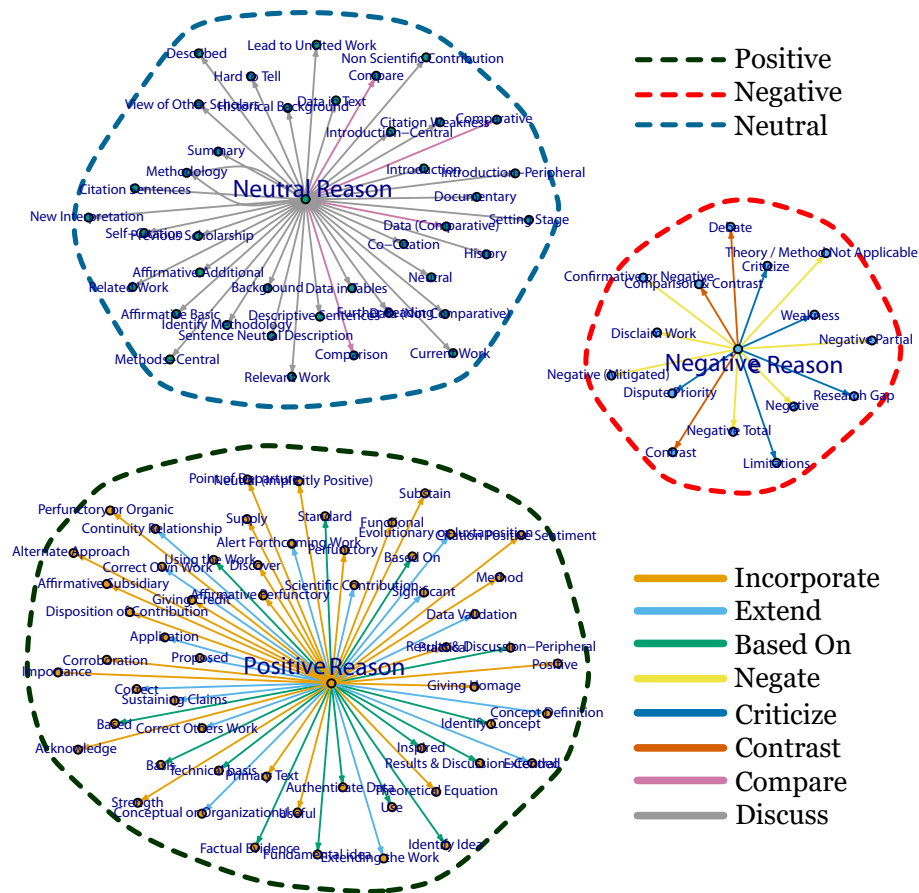


FIGURE 4.2: Citation Reason Clusters and Properties.

4.3 Formal Definition of Classes and Properties

Based on the experiments and their results, we have defined “Citation’s Context and Reasons Ontology – CCRO”, both at the schema and instance level. CCRO defines a taxonomical hierarchy of eight object properties distributed among three main sentiment-based reasons. Using the ontology concept, a citation can have “Positive”, “Negative” or “Neutral” context or sentiment with the possibility of being a part of one of its constituent properties. Fig 4.3 describes classes and object-properties for Ontology. The name-space used is “*http://ccropus*” for *CCRO* and is abbreviated as *ccro*. The ontology is defined at both instance and schema level. CCRO has been verified, validated, and assessed by using well-defined procedures and tools proposed in the literature for ontology evaluation (Section: 4.4). The results show that the proposed ontology is concise, complete, and consistent. For the instantiation and mapping of ontology classes on real data, we have developed a *Mapping Graph* among the verbs with predicative complements in the English Language, the verbs extracted from the selected corpus using NLP and CCRO properties. Using this *Mapping Graph*, mapping of ontology classes in each citation’s sentiment is explained, with a complete mapping of the selected corpus.

4.3.1 CCRO Classes

The base class of *CCRO* is a “*Paper*” that corresponds to any type of research documents such as a book, conference paper, journal article, presentation, report, or thesis, etc. It has two subclasses “*ccro:CitingPaper*” and “*ccro:CitedPaper*”. “*ccro:CitingPaper*” is any document that refers to another document whereas the document being referred is the “*ccro:CitedPaper*”. Each citing paper consists of citation texts that are structured around the main verbs. Two classes “*ccro:Citation*” and “*ccro:MainVerb*” define this concept. “*ccro:MainVerb*” refers to the words with part of speech as a verb and is equivalent to the class of the main verb in *OLiA*¹ ontology, an annotation model based on morphology.

¹<http://nachhalt.sfb632.uni-potsdam.de/owl/olia.owl>

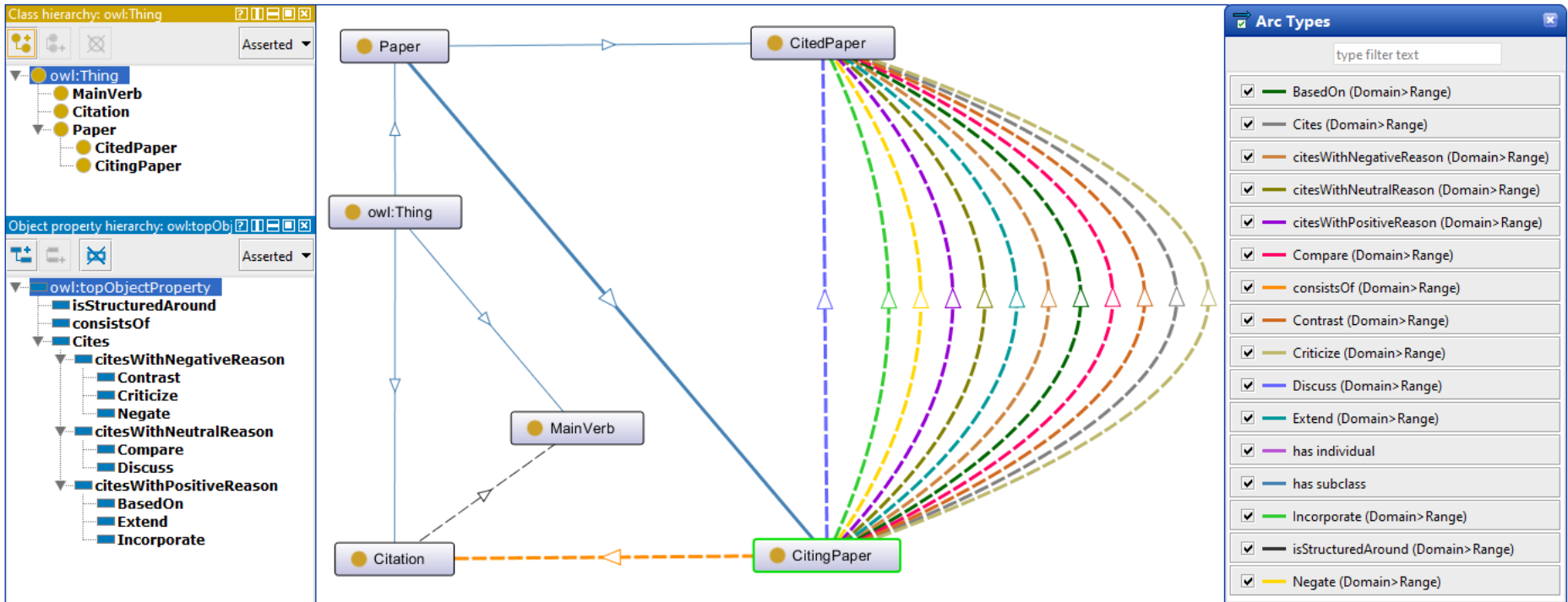


FIGURE 4.3: Citation's Context and Reasons Ontology

4.3.2 CCRO Object Properties

CCRO defines Object Properties in a hierarchy with three main properties that are “*ccro:consistsOf*”, “*ccro:isStructuredAround*” and “*ccro:Cites*”. “*ccro:consistsOf*” defines citation text with domain “*ccro:CitingPaper*” and range “*ccro:Citation*”. As defined earlier, each citation-text is structured around a main verb, thus “*ccro:isStructuredAround*” is used for the said purpose with a domain “*ccro:Citation*” and range “*ccro:MainVerb*”.

The third property is a hierarchy of properties that starts with a coarsely defined citation reason and then incrementally, it is further defined at each level. All properties in the hierarchy have similar domain and range that is “*ccro:CitingPaper*” and “*ccro:CitedPaper*” respectively. The hierarchy has three levels. At the base-level is the property “*ccro:Cite*” that defines any generic reason for citation. The next level decomposes the nature of citations based on sentiment with three classes that are:

1. “*ccro:citesWithPositiveReason*”
2. “*ccro:citesWithNegativeReason*”
3. “*ccro:citesWithNeutralReason*”

All the above three properties are disjoint whereas positive and negative reason classes are inverse of each other. The third level defines cognitive reason property based on our experiments. These properties are distributed among different sentiments and refer to a specific type of reason that is associated with an English verb category in the dictionary with shared meanings and syntactical behavior. There are eight properties, first three are sub-properties of positive, the next three are of negative and the last two are of neutral reasons. These properties are:

1. “*ccro:Incorporate*”
2. “*ccro:Extend*”

3. “*ccro:BasedOn*”
4. “*ccro:Negate*”
5. “*ccro:Criticize*”
6. “*ccro:Contrast*”
7. “*ccro:Compare*”
8. “*ccro:Discuss*”

4.3.3 CCRO Instance Level Schema

After defining the schema level classes and properties, the next step is to encode a mapping between the extracted individual English verbs and the corresponding reason classes using the dictionary. We instantiate each English verb at the instance level with a “*ccro:MainVerb*” class and then map the verb with associated reason property. For example, two of the most common verbs used in ACL Anthology Network Citation Data are “Use” and “Describe”. These two verbs will be initially instantiated as “*ccro:MainVerb*”. Next, using the sentiment polarity of the citation in which these verbs are used, a mapping connects the citations to their corresponding reason classes. It should be noted here that each individual verb can be associated with several reasons based on the context (sentiment) in which that verb is used.

Definition: Verb “V”: “*V verb is an instance of “ccro:MainVerb” class and refers to ‘any reason for citing’ based on the sentiment of the citation text in which it is used. ”*

A small example, shown in Fig 4.2, describes two verbs when used in different sentiments, with reference to their corresponding classes in a graphical representation.

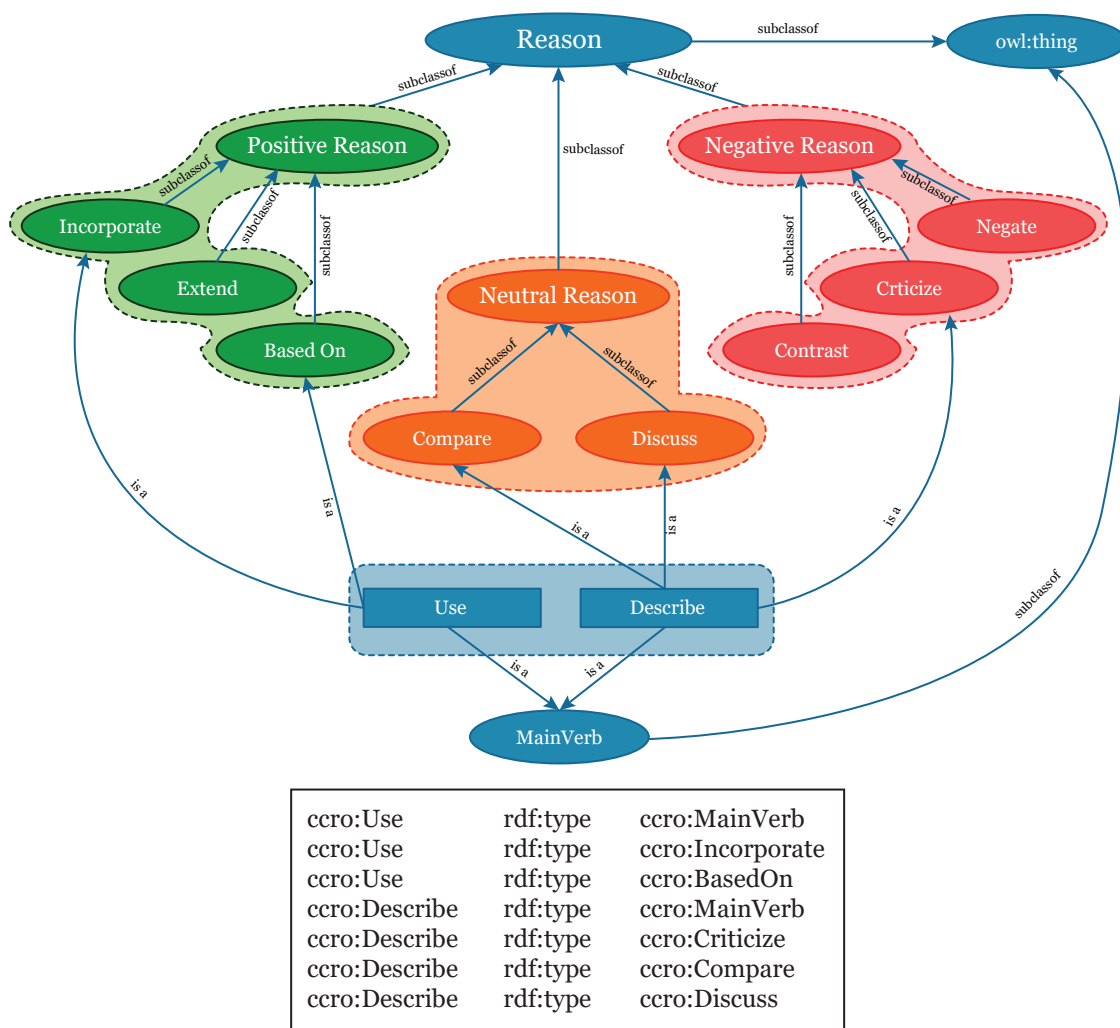


FIGURE 4.4: CCRO - Instance Level Schema Definition

4.3.4 CCRO Citation Graph

Scientific literature can be visualized as a networked information space that contains a collection of information-entities inter-connected by a link-structure. In this interconnected network, scientific papers are “information entities” and their links are defined as citations. A citation graph represents the network where each paper becomes the node and their citation link becomes the edge, between the nodes. The citation link also implies that the citation graph is a directed graph where one paper cites the paper, at the other node. *CCRO* classes define the edge between these nodes making the citation graph a semantic graph where each edge is represented using an *RDF* triple as shown in Fig 4.5.

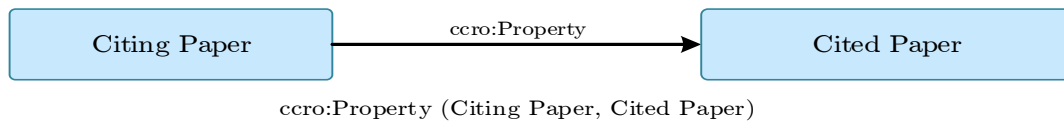


FIGURE 4.5: CCRO - Citation Link RDF Triple Graph

TABLE 4.3: CCRO Evaluation using Automated Tools

Tool Name	Tool Evaluation	Evaluation Result
OWL Validator	Checking OWL2 Syntax	No errors detected
Hermit Reasoner	Object & Data Properties	No errors detected
Fact++ Reasoner	Inconsistency Errors	No errors detected

4.4 Ontology Evaluation and Validation

Ontologies can be evaluated on their syntactic and semantic structure, normally referred to as content-based evaluation. There are two methods available to evaluate an ontology, one is using automated evaluation tools and the other is by user-study. To evaluate the proposed ontology, both approaches are used.

4.4.1 Automated Tool Evaluation

Various automated evaluation tools can evaluate the contents of the ontology for checking the syntax of language used to describe the ontology. In our experiment, we have used *OWL2 Validator*² for property analysis, *Hermit Reasoner* [106] to identify basic errors in the ontology and *Fact++ Reasoner*³ for inconsistency errors. Both Hermit and Fact++ Reasoners are configured on Class, Object Property, Data Property, and Individual preferences. Figure 4.6 represents both reasoners used in the Protégé environment. Findings of each tool are shown in Table 4.3.

²<http://mowl-power.cs.man.ac.uk>

³<http://owl.man.ac.uk/factplusplus/>

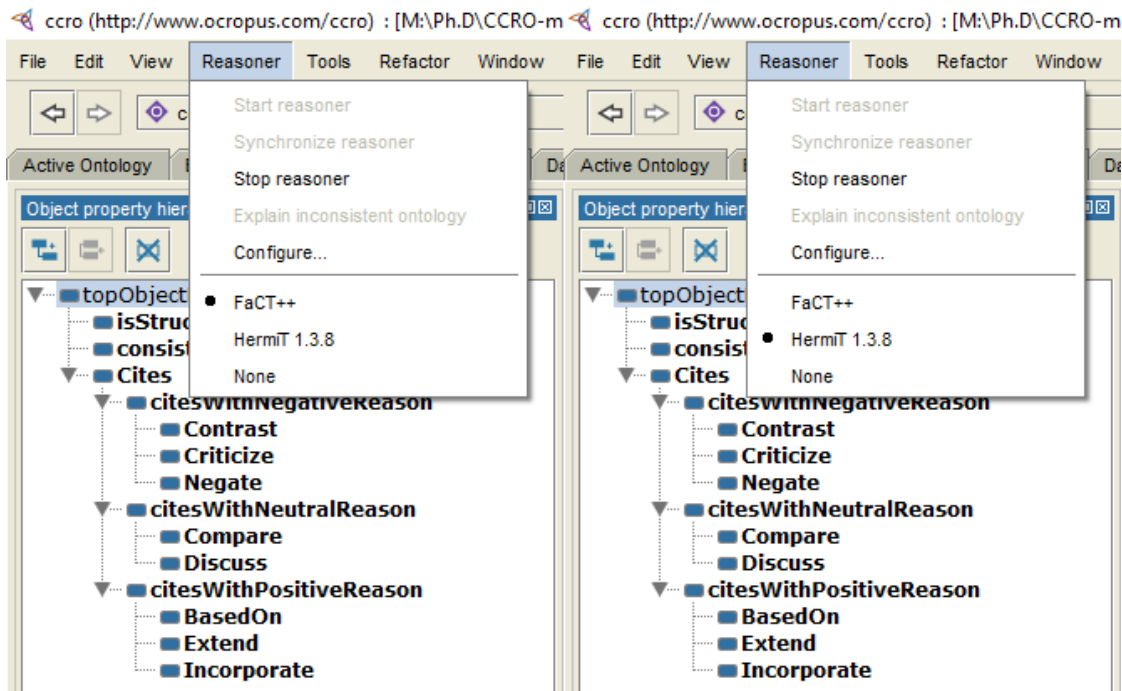


FIGURE 4.6: Fact++ and Hermit Reasoners for CCRO

4.4.2 User-Based Study Evaluation

To evaluate the proposed ontology on a user study, a questionnaire was designed using the outline provided by [79, 107]. The questionnaire deals with “Incomplete Concept Classification”, “Disjoint Knowledge Omission”, “Exhaustive Knowledge Omission”, “Scientific Knowledge Omission” and “Redundancy of Disjoint Relations”. Each area contains two questions. First, a binary “Yes or No” question asking if the issue exists or not, and second is a descriptive question based on the answer of the first. The questions and their objectives are outlined in Appendix 10. Ten (10) domain experts with vast experience in Digital Library, NLP, and Computational Linguistics evaluated the ontology using the prescribed questionnaire. Profiles of selected evaluators are shown in Table 4.4.

The results of the evaluation are shown in Fig 4.7 whereas Table 4.5 outlines omissions and recommendations suggested by the experts.

TABLE 4.4: CCRO’s Evaluators Profiles

Sr.	Evaluator Profile	Evaluators
1.	Professor (Semantic Computing & Ontology Engineering)	1
2.	Associate Professor (NLP and Machine Learning)	1
3.	Associate Professor (Corpus Linguistics)	1
4.	Assistant Professor (Corpus Linguistics)	1
5.	PhD Students (Semantic Computing & Ontology Engineering)	2
6.	PhD Students (Corpus Linguistics)	2
7.	Lecturers (English Language and Linguistics)	2

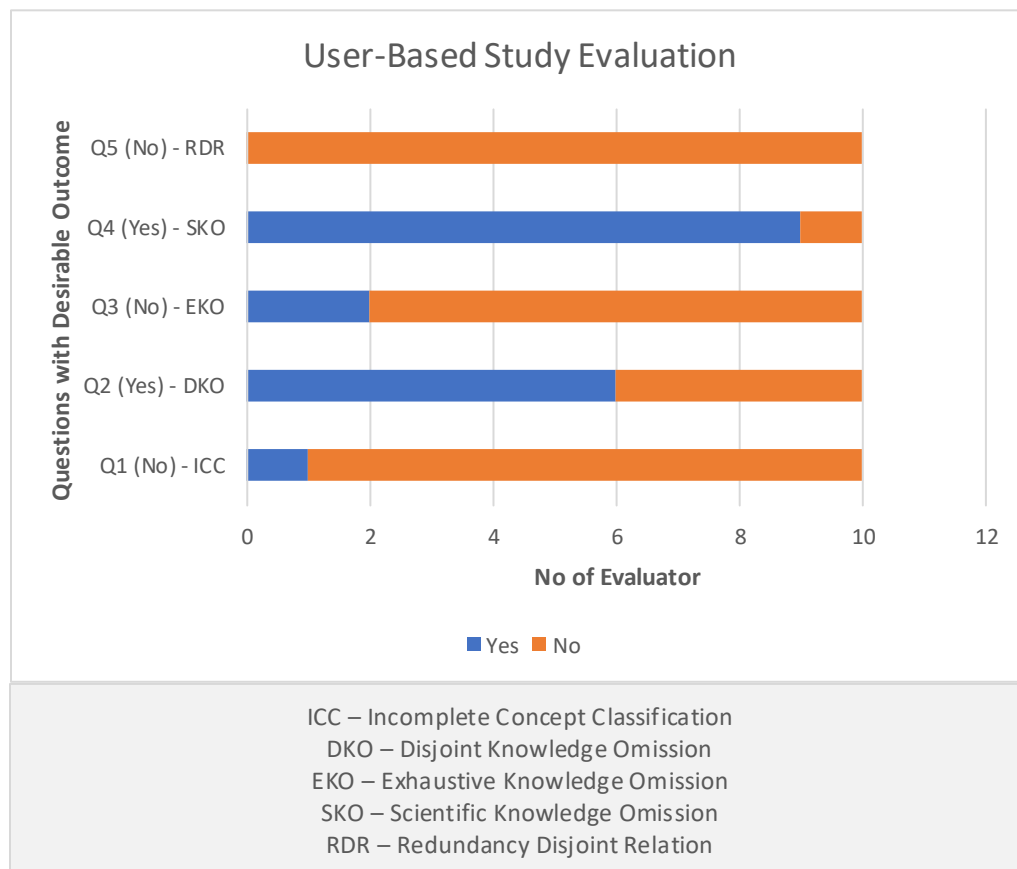


FIGURE 4.7: User Study Evaluation Report for CCRO

TABLE 4.5: CCRO Evaluation by Domain Experts

Error Type	Yes or No	Omissions / Recommendations
Incomplete Concept Classification	No	All concepts included.
Disjoint Knowledge Omission	Yes	“Positive” and “Negative” Citation Reasons are naturally disjoint. However, these two are NOT disjoint with “Neutral” Citation Reasons. Moreover, each citation reason property within “Positive”, “Negative” and “Neutral” Reasons might not be disjoint.
Exhaustive Knowledge Omission	Yes	Publish Ontology Online
Sufficient Knowledge Omission	Yes	All concepts defined properly
Redundancy Disjoint Relation	No	No Redundancy in Disjoint Relations found

4.4.3 Findings and Results

CCRO consists of eight disjoint classes that are grouped into “Positive”, “Negative” and “Neutral” reasons. However, user-based evaluation has pointed out a critical point in the domain of “Disjoint Knowledge Omission”. The study suggests that “Positive” and “Negative” Citation Reasons are naturally disjoint but it does not mean that all three “Positive”, “Negative” and “Neutral” are disjoint. By definition, two sets are disjoint if they have no elements in common. However, authors can choose to cite research with “Positive” and “Neutral” or “Negative” and “Neutral” context. Furthermore, the study has revealed that *CCRO* Properties within each sentiment context can also have overlapped meanings. Thus one citation can be mapped in multiple classes within a sentiment context. Based on the user-based study evaluation and recommendations, updated *CCRO* properties are shown in Fig 4.8.

As discussed earlier, there is a difference between a citation and a reference [14]. The reference refers to the works mentioned in the reference section or bibliography of a journal article. A reference may be mentioned once or multiple times in an

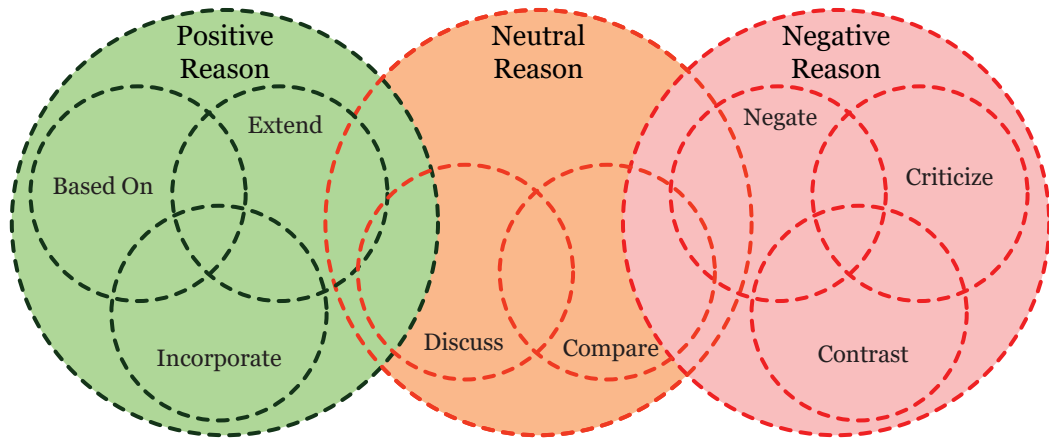


FIGURE 4.8: CCRO Properties Based on User-Based Study Evaluation

article. Each mention is considered as a citation. Thus, there is a possibility that an author cites the same article in “Positive” and “Neutral” or “Negative” and “Neutral” sentiment contexts. However, it is very difficult that the same article is cited in “Positive” and “Negative” sentiment contexts in the same article. Therefore, a rule is integrated within *CCRO* to enforce this. The rule is:

“ccro:citesWithPositiveReason owl:disjointWith ccro:citesWithNegativeReason”

This is the only disjoint rule within *CCRO*.

Three sub-properties in *ccro:citesWithPositiveReason* will automatically become disjoint with three sub-properties in *ccro:citesWithNegativeReason*. Using a simple example, Fig 4.9 shows two citations with “Negative” and “Neutral” sentiments for the same citing and cited paper. Therefore, we can say that even though Citation Reasons have disjoint nature, their instances can have an overlapping behavior. Based on these findings, an updated ontology is now publicly available at <https://github.com/imranihsan/CCRO/blob/master/CCRO.owl>.

4.5 Conclusion

A scientific research paper contains vital information that incites its citation by the authors and researchers based on diverse and innumerable reasons. After a thorough survey on semantic representations and limitations of these reasons,

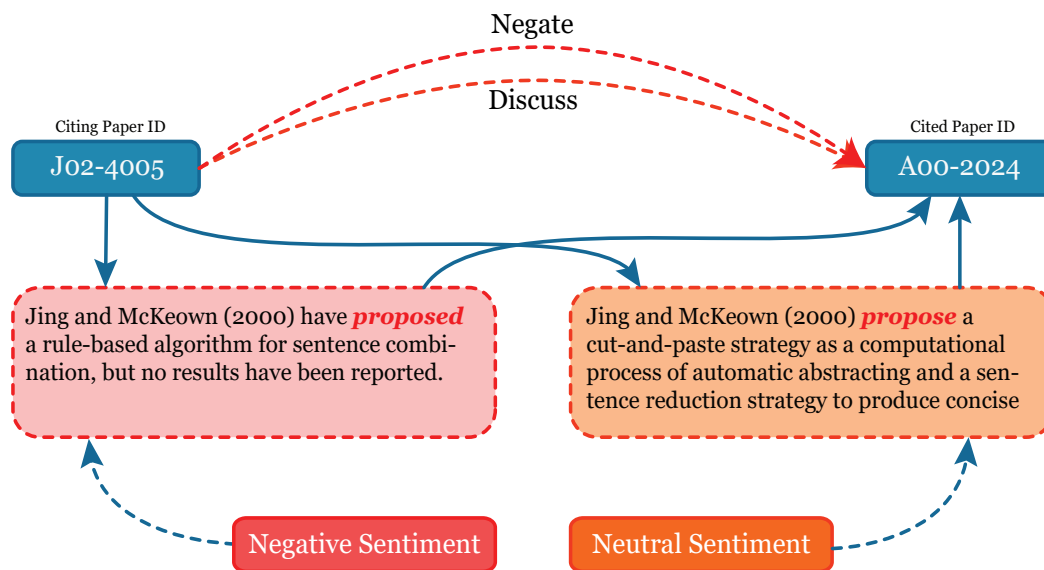


FIGURE 4.9: Example of “Negative” and “Neutral” Sentiment Citations for Same Citing and Cited Papers

an ontology, titled “CCRO: Citation’s Context and Reasons Ontology”, is being proposed that provides the abstract conceptualization required to organize the citations’ relations. This ontology successfully distills down into eight distinct and mutually exclusive classes, all routed from “Positive”, “Negative” and “Neutral” sentiments. The available tools and techniques to evaluate and validate an ontology are also applied to make the proposed ontology concise, complete, and consistent.

Chapter 5

Ontology Instantiation and Mapping

After the development and evaluation of proposed ontology, the next step is to formulate experiments to instantiate ontology classes and properties on real data. Using the methodology for ontology instantiation and mapping (Discussed in Section 3.2), two more experiments are performed. Both use citation corpora *CC1* and *CC2*. As described in methodology, *CC1* is publicly available ACL Anthology Network - AAN Data-set with 7390 citation texts and *CC2* is our own manually created data-set with 1230 citation texts. The first experiment (Section 5.1) is pre-processing of citation texts to extract *Reporting Verbs* by identifying unique verbs from selected corpora with their frequencies, and their distribution across all three sentiments [108]. With the help of Levin’s knowledge base [81], a *Mapping Graph* is formulated among the extracted unique Verbs, Levin’s verbs with predicative complements and *CCRO* Reason properties. The second experiment (Section 5.2) is automatic annotation of selected corpora on *CCRO* reason properties using the *Mapping Graph*. The last experiment (Section 5.3) is to manually annotate each citation text in the selected corpora on *CCRO* reason property with help of experts and annotators. The resultant will be *CCRO* distribution on both corpora *CC1* and *CC2* (Section 5.4). The complete process is shown in Fig 5.1.

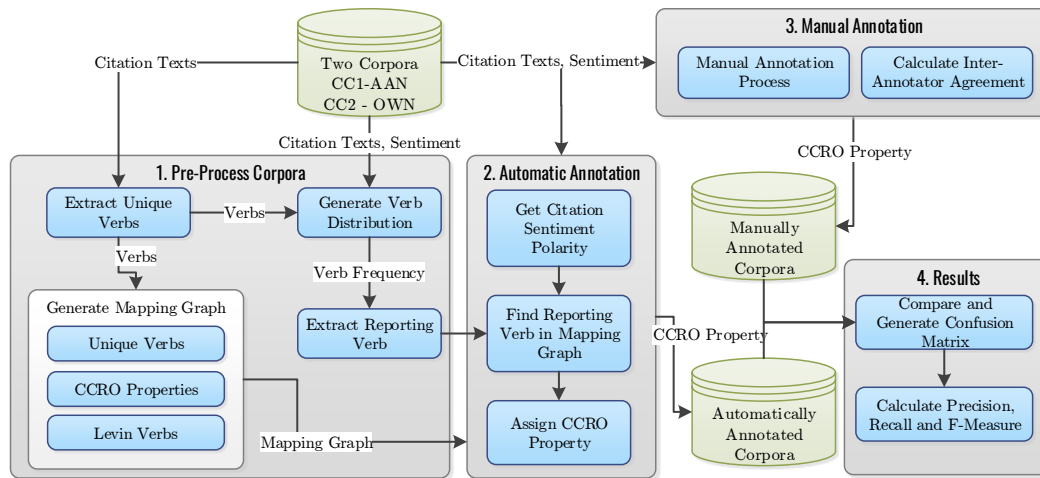


FIGURE 5.1: Methodology for CCRO Instantiation and Mapping

5.1 Pre-Process Corpora

5.1.1 Extract Unique Verbs

The POS Tagger can read a text string and assign parts of speech to each word such as *Noun*, *Verb*, *Adjective* etc. We have developed an application that reads corpus *CC1* line by line. Each line contains a citation text. POS Tagger tags verbs in the citation text in six different forms.

1. **VB Verb**, base form
2. **VBD Verb**, past tense
3. **VBG Verb**, gerund or present participle
4. **VBN Verb**, past participle
5. **VBP Verb**, non3rd person singular present
6. **VBZ Verb**, 3rd person singular present

Once the tagging is complete, lemmatization and stemming are applied alternatively to convert the verbs in its basic form. Afterward, a simple algorithm finds

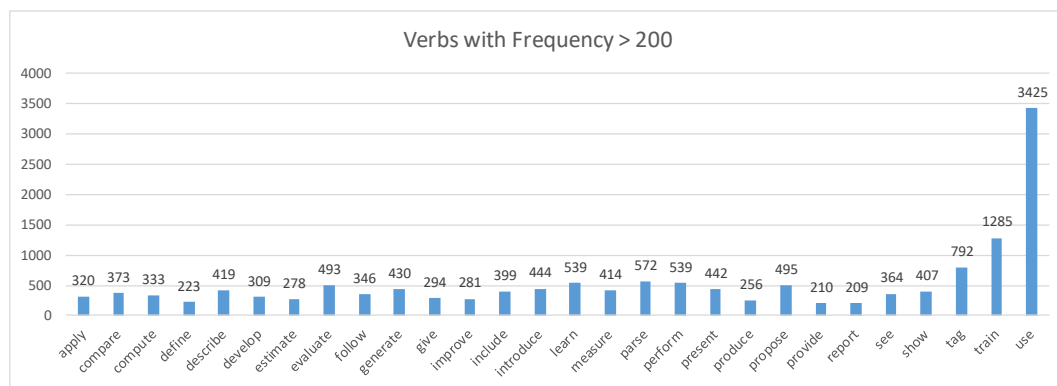


FIGURE 5.2: Unique Verbs with Frequency ≥ 200 in AAN Corpus

the unique verbs and calculates their frequency (No of times used in the complete corpus). The results show that there are 389 unique verbs present. To find the accuracy of results, we have randomly selected five sets of 100 citation texts each and have asked our language experts to mark the verbs. After careful comparison of manual and automated extracted verbs, a confusion matrix is generated for “True Positive”, “True Negative”, “False Positive” and “False Negative”. The experiment reveals that the accuracy of the automated verb extraction application is 88.25%. Results also suggest that the verb “Use” is the most common verb used by the authors with a frequency of 3425. The second most common verb is “Train” with a frequency of 1285 and “Tag” becomes the third most common verb with a frequency of 792. Fig 5.2 shows verbs with a frequency greater than 200 within the AAN corpus.

After finding the frequency of verbs in the complete corpus, the next step is to analyze the distribution of these verbs in all three sentiments. Using the sentiment polarity information available in the selected corpus, each verb frequency is checked in citation texts with “Positive”, “Negative” and “Neutral” sentiments. The corpus contains 829 citation texts with “Positive” sentiment, 280 with “Negative” and 7627 with “Neutral” sentiment. The frequency percentage of i^{th} verb in “Positive” sentiment F_p^i is calculated using the formula shown in Eq 1. Similarly, frequency percentages of i^{th} verbs in “Negative” F_n^i (Eq 2) and “Neutral” F_o^i (Eq 3) sentiments can also be calculated. The resultant distribution for the top 20

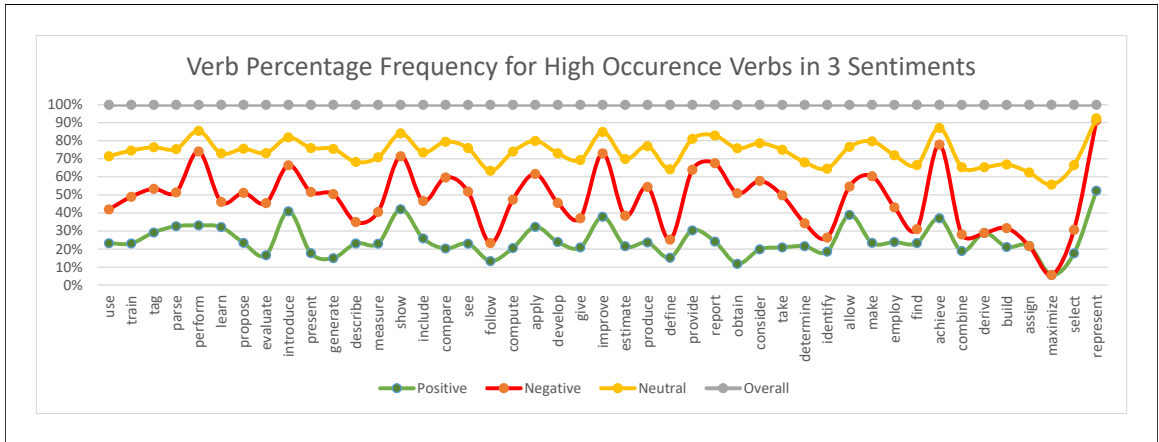


FIGURE 5.3: High Occurrence Verbs Distribution in 3 Sentiments

verbs is shown in Fig 5.3. Table 5.1 describes a selection of verbs and no of the times, these verbs have been used by various authors in each sentiment.

$$F_p^i = \frac{\sum V_p^i * 100}{\sum V_p^i + \sum V_n^i + \sum V_o^i} \quad (5.1)$$

$$F_n^i = \frac{\sum V_n^i * 100}{\sum V_p^i + \sum V_n^i + \sum V_o^i} \quad (5.2)$$

$$F_o^i = \frac{\sum V_o^i * 100}{\sum V_p^i + \sum V_n^i + \sum V_o^i} \quad (5.3)$$

where V_p^i describes i^{th} verb used in citation text with “Positive”, V_n^i describes i^{th} verb used in citation text with “Negative” and V_o^i describes i^{th} verb used in citation text with “Neutral” sentiment.

Semantic-based analysis of the selected corpus shows that the use of verbs is a crucial lexical element for authors to cite others and claim their findings. Frequency percentage distribution of verbs in all three sentiments exhibit correlation, however, some of the verbs show deviant behavior, with a higher number of percentages in “Positive” or, “Negative” but lower in “Neutral” sentiment, as compared to other verbs. In general, verbs appearing in citation-texts with “Positive” or “Negative” sentiment are more assertive, rather than the verbs that appear in citation texts with “Neutral” sentiment. Let’s investigate all three sentiments in detail.

TABLE 5.1: Verbs and No of Times Used in All 3 Sentiments

Verb	No of Times Used in Sentiment			Total
	Positive	Negative	Neutral	
Use	264	72	3089	3425
Train	111	42	1132	1285
Tag	93	26	673	792
Parse	72	14	486	572
Perform	117	49	373	539
Learn	61	9	469	539
Propose	45	18	432	495
Evaluate	29	17	447	493
Introduce	95	20	329	444
Present	31	20	391	442
Generate	25	20	385	430
Describe	29	5	385	419
Measure	31	8	375	414
Show	102	24	281	407
Include	37	10	352	399
Compare	35	23	315	373
See	33	14	317	364
Follow	12	3	331	346
Compute	25	11	297	333
Apply	49	15	256	320
Develop	26	8	275	309

TABLE 5.1: Verbs and No of Times Used in All 3 Sentiments (Continued)

Verb	No of Times Used in Sentiment			Total
	Positive	Negative	Neutral	
Give	19	5	270	294
Improve	67	21	193	281
Estimate	19	5	254	278
Produce	25	11	220	256
Define	9	2	212	223
Provide	32	12	166	210
Report	28	17	164	209
Obtain	9	10	174	193
Consider	17	11	165	193
Take	15	7	167	189
Annotate	16	4	154	174
Classify	17	0	141	158
Determine	10	2	144	156
Identify	7	1	133	141
Allow	22	3	114	139
Make	15	8	114	137
Consist	11	4	122	137
Employ	11	3	122	136
Find	9	1	126	136
Achieve	35	13	80	128
Combine	6	1	109	116
Contain	8	2	101	111

5.1.1.1 Distribution of Verbs in Positive Sentiment

Distribution of verbs in “Positive” sentiment shows that the top 5 frequent verbs are “Use”, “Perform”, “Train”, “Show” and “Introduce”. A common behavior of these verbs shows that authors use these verbs to incorporate research as a part of a whole, extend their research from a central point to a wider solution, or base their research on the foundation of other researches, when used in a “Positive” sentiment within an academic paper.

5.1.1.2 Distribution of Verbs in Negative Sentiment

Distribution of verbs in “Negative” sentiment shows that the top 5 frequent verbs are “Use”, “Perform”, “Train”, “Tag” and “Show”. A common behavior of these verbs shows that authors use these verbs to negate research, criticize and find fault in research or provide a contrast showing differences with opposite nature when used in a “Negative” sentiment within an academic paper.

5.1.1.3 Distribution of Verbs in Neutral Sentiment

Distribution of verbs in “Neutral” sentiment shows that the top 5 frequent verbs are “Use”, “Train”, “Tag”, “Parse” and “Learn”. A common behavior of these verbs shows that authors use these verbs to compare techniques by showing similarities or simply discuss and examine by argument when used in a “Neutral” sentiment within an academic paper.

5.1.2 Extract Reporting Verb

Reporting Verbs are one of the crucial components in academic writing. Many types of research have been conducted in past to analyze the *Reporting Verbs* in doctoral theses, students’ assignments, research articles, and journals [93]. With the help of *Reporting Verbs*, authors use the most suitable word to relate the source

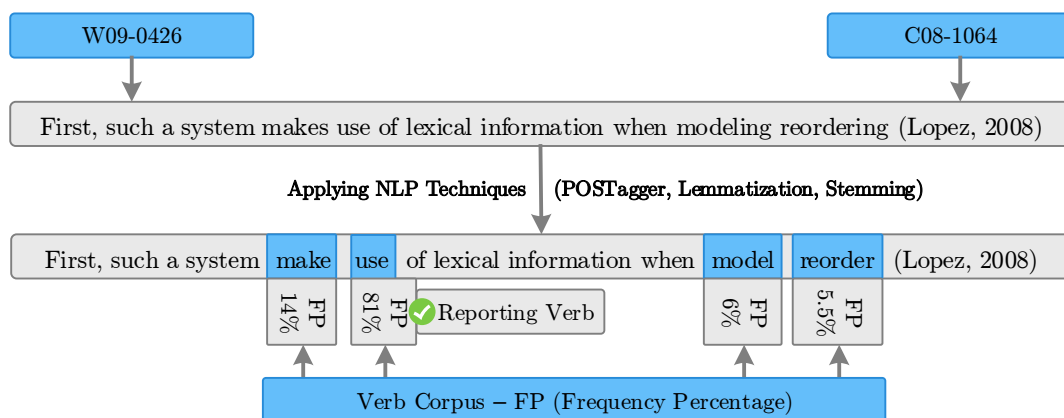


FIGURE 5.4: Extract Reporting Verb from Citation Text using NLP and Verb Corpus

which they have found convincing and suitable to support their claim and can become the basis of citation reason [109]. The study also suggests that different authors mostly used similar *Reporting Verbs* [93]. Therefore, the frequency of a verb can play a vital role in identifying a *Reporting Verbs*. Verbs with high frequencies normally mean that they have been commonly used to cite a research article by different authors. Therefore, if a citation text contains more than one verb then the verb with the highest frequency will act as the *Reporting Verb*. An application has been developed using *Python NLP* to extract *Reporting Verb* using formula show in Eq 5.4 and its working in Fig 5.4.

$$V_R = \max_{\{V_i\}} [F_p^i + F_n^i + F_o^i] \quad (5.4)$$

where V_R is the *Reporting Verb*.

5.1.3 Generate Mapping Graph

Levin’s lexical knowledge [81] defines each English verb class using two characteristics: “semantically coherent” and “shared syntactic behavior”. Semantically coherent means the verbs exhibit unique properties that have shared meanings. An individual verb can be associated with multiple classes depending upon the

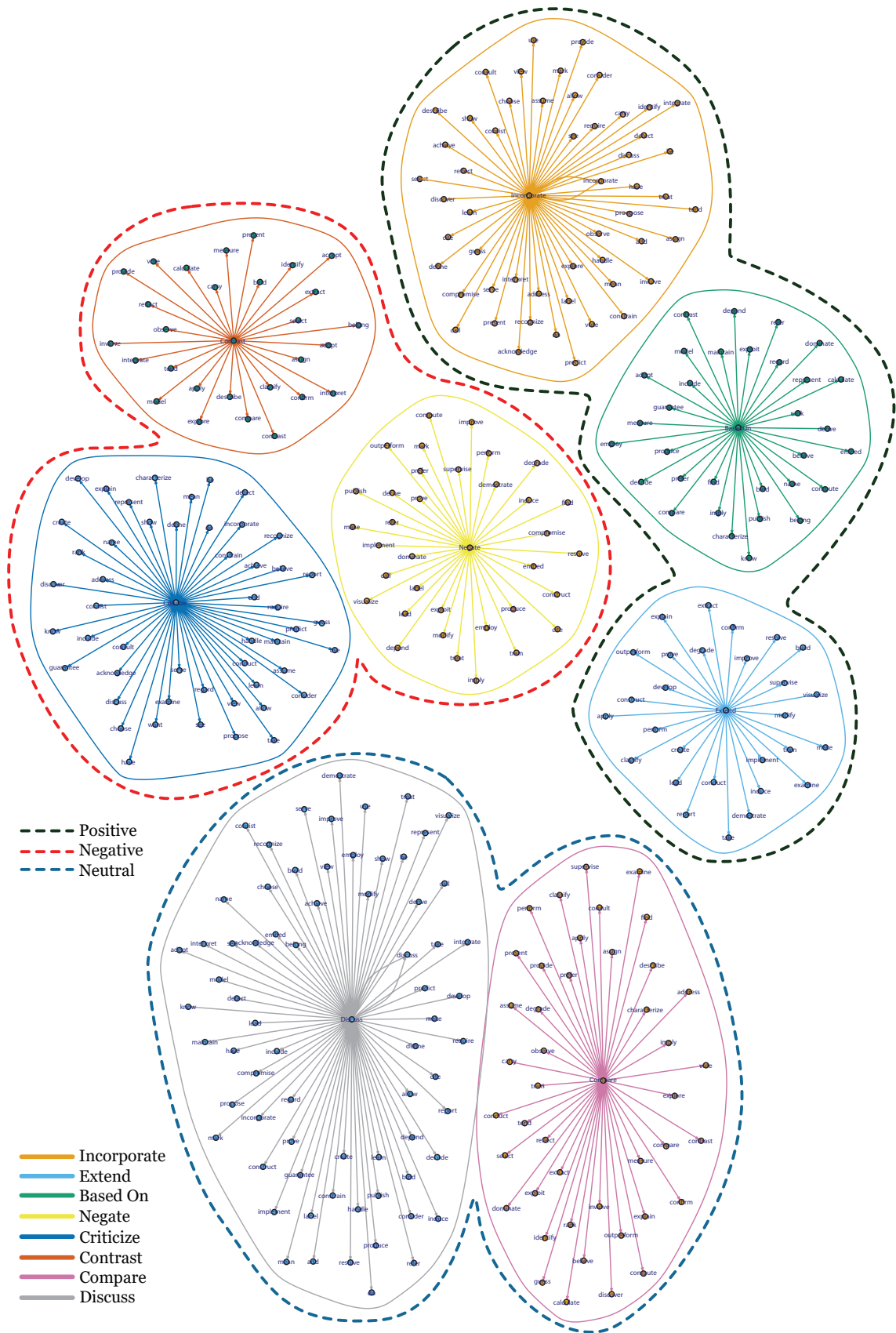


FIGURE 5.5: Partial Representation of Dictionary – A Mapping between Levin Conceptual Hierarchy Verbs and Citation Reason Classes

context in which it is used, thus making the relationship a graph rather than a tree. Whereas shared syntactic behavior describes a verb in terms of expression and interpretation. Verbs that share meaning exhibit similar syntactic behavior [110].

Using this conceptual hierarchy of verb classes, we have defined a *Mapping Graph*. The unique properties and collaborative meaning of each verb class correspond to each citation class that has similar collaborative properties. Each verb can have “Positive”, “Negative” or “Neutral” sentiment based on its context. Therefore, we have visualized each verb in all three sentiments making them semantically coherent. A complete mapping of the classified verbs based on Levin Conceptual Hierarchy and Citation Reason Properties formulates the *Mapping Graph*. A group of four English Linguistics experts (2 Lecturers and 2 MS Students) have helped in formulating this dictionary. Figure 5.5 represents a partial representation of the *Mapping Graph*.

5.2 Automatic Mapping of Citations’ Reasons

After extracting *Reporting Verbs* and *Sentiment Polarity* of a citation text, a mapping model is required to map *Reporting Verb* on a *CCRO* property. Working on this model is shown in Figure 5.6. *Mapping Graph* contains a list of verbs against each *CCRO* property. Therefore, we have developed a small application in Python that reads each citation text from both corpora (*CC1* and *CC2*) along with its sentiment polarity and extracts Reporting Verb. With a help of *Mapping Graph*, the selected Reporting Verbs is mapped onto one or more *CCRO* properties, formulating a cognitive relationship between citing and cited paper. A complete mapping procedure in all 3 sentiments is shown next. To create an instance of *CCRO*, two additional vocabularies are used to annotate citations. These vocabularies are NIF¹ (NLP Interchange Format) and BiRO² (Bibliographic Reference Ontology)

¹<http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html>

²<http://www.sparontologies.net/ontologies/ biro/source.html>

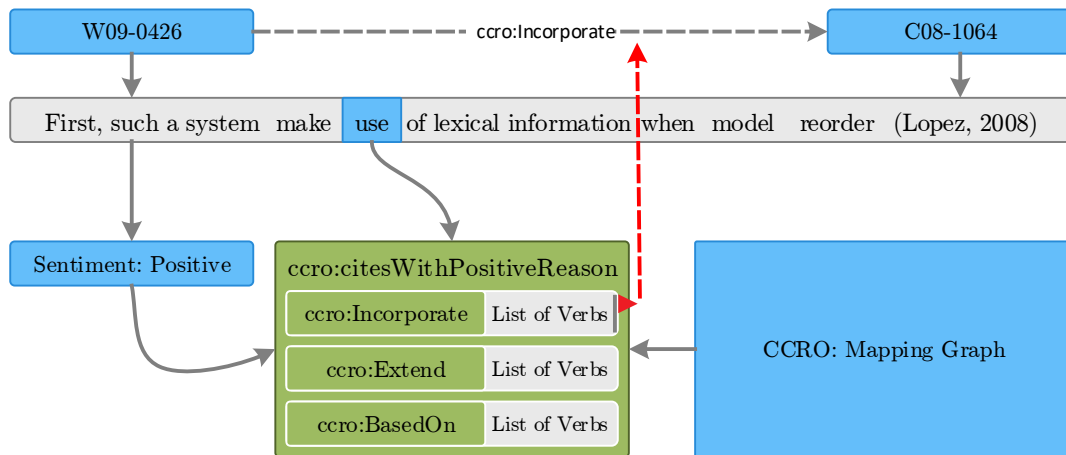


FIGURE 5.6: Mapping Model to Map Reporting Verb on CCRO Property using Mapping Graph and Sentiment Value

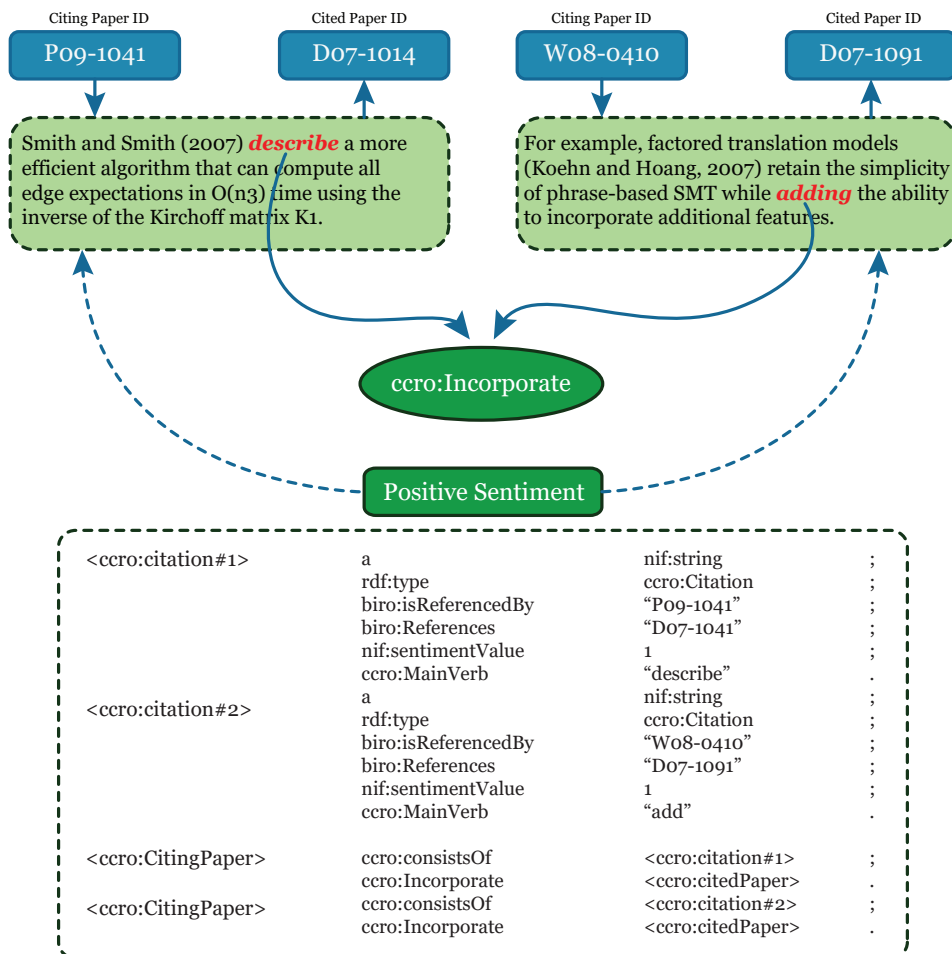


FIGURE 5.7: Mapping Citation with Positive Sentiment on CCRO

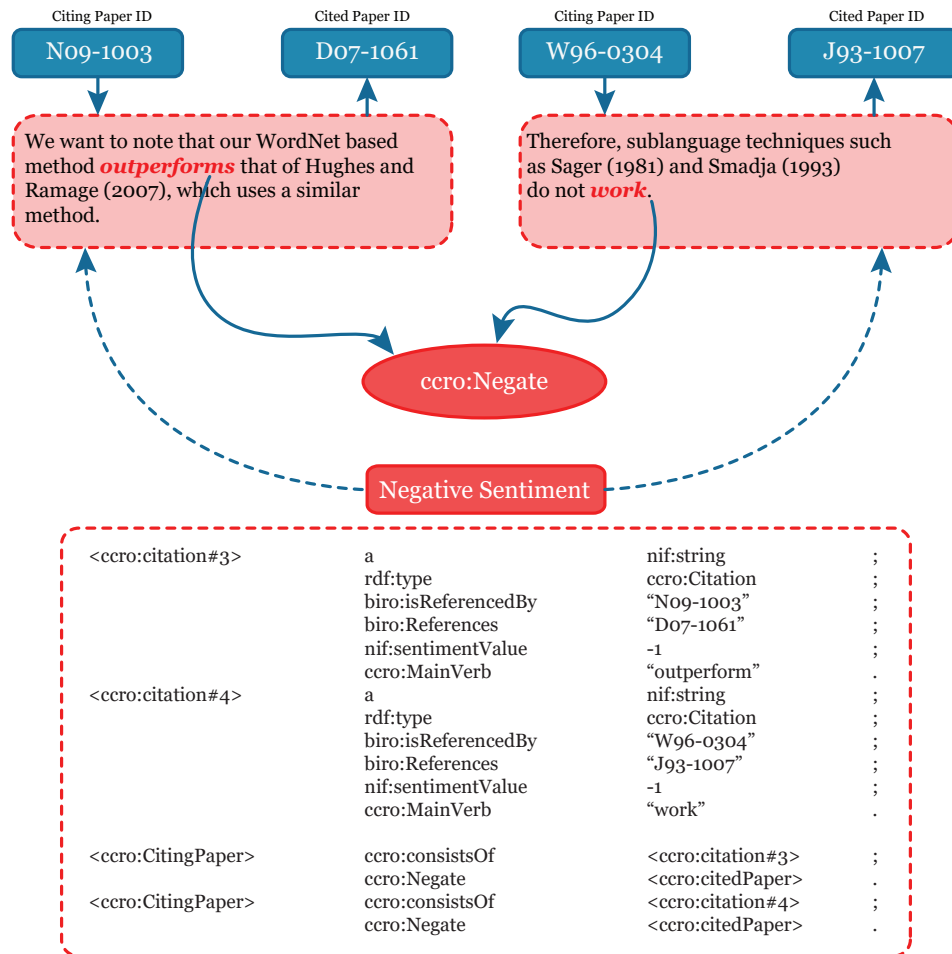


FIGURE 5.8: Mapping Citation with Negative Sentiment on CCRO

5.2.1 Mapping Citation with Positive Sentiment on CCRO

Figure 5.7 shows two citation texts with a Positive sentiment. The first ID is the citing paper and the second ID is the cited paper. **Citation#1** is structured around the verb “describe” and **Citation#2** around the verb “add”. Both the citations share a similar meaning in the sense that both are incorporating other algorithms in their research. Using the *Mapping Graph*, annotating these two citation texts via *CCRO* leads us to obtain the reason as “ccro:Incorporate”.

5.2.2 Mapping Citation with Negative Sentiment on CCRO

Figure 5.8 shows two citation texts with a Negative sentiment. **Citation#3** is structured around the verb “outperform” and **Citation#4** around the verb

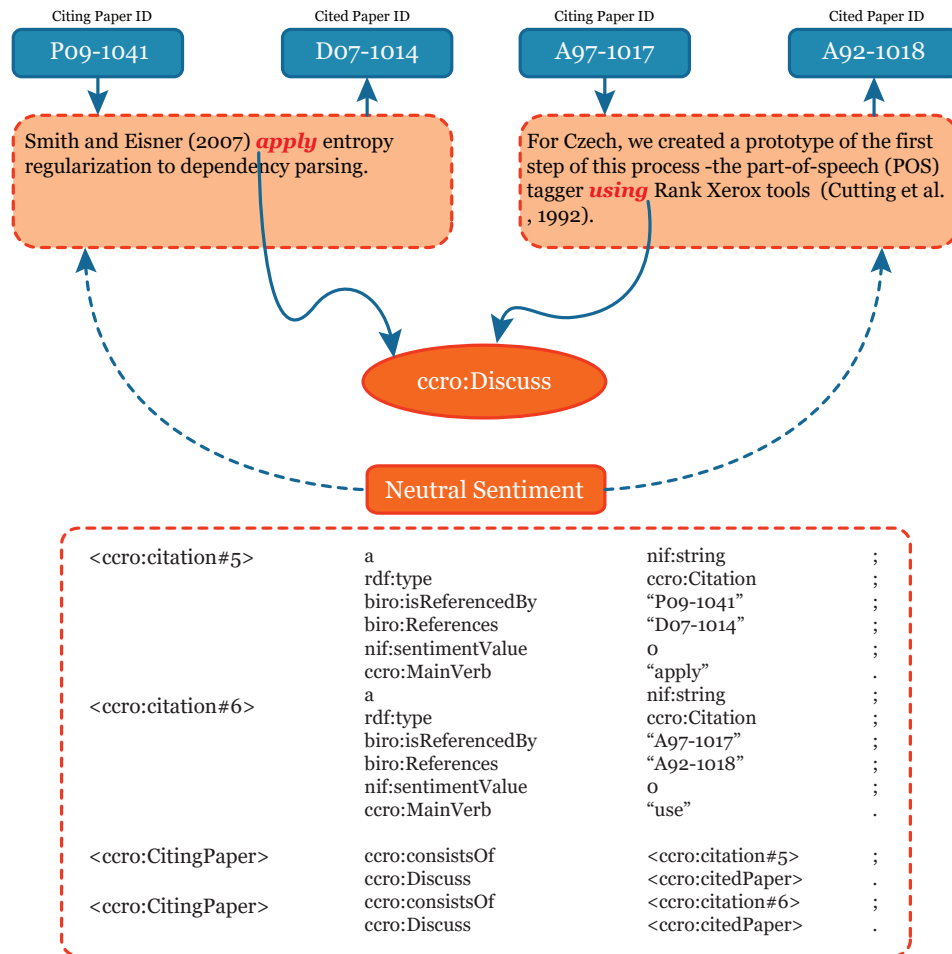


FIGURE 5.9: Mapping Citation with Neutral Sentiment on CCRO

“work”. Both the citations share a similar meaning in the sense that both are negating existing work. Using the *Mapping Graph*, annotating these two citation texts via *CCRO* generates “ccro:Negate”.

5.2.3 Mapping Citation with Neutral Sentiment on CCRO

Figure 5.9 shows two citation texts with a Neutral sentiment. Citation#5 is structured around the verb “apply” and Citation#6 around the verb “use”. Both the citations share a similar meaning in the sense that both are discussing existing works as a literature review. Using the *Mapping Graph*, annotating these two citation texts via *CCRO* leads us to obtain “ccro:Discuss”.

TABLE 5.2: CCRO’s Annotators Profiles

Sr.	Annotator Profile	Annotators
1.	MS/PhD Students (Scientometrics & H-Index)	2
2.	MS/PhD Student (Semantic Computing & Ontology Engineering)	1
3.	MS/PhD Student (Data Science & Sentiment Analysis)	1
4.	MS/PhD Students (Corpus Linguistics)	2

5.3 Manual Annotation of Citations’ Reasons

To find the accuracy of automated annotation of citation texts with citations’ reasons, manual annotation of both corpora is performed. The manual annotation process requires a collaborative approach between the domain experts and the designers of the annotation process. This is a natural language processing job. Therefore it can have problems and can lead to no consensus [111]. To achieve maximum inter-annotator agreement, six annotators are selected, four of them are domain experts and two of them are linguists. Profiles of selected annotators are shown in Table 5.2. To annotate corpora *CC1* and *CC2* on *CCRO* properties, an annotation tool is developed that can enable annotators to view and annotate a citation on citation reason class with ease. Corpus *CC2* has four domains with 10 papers each. Therefore, four experts are provided with 10 domain and 10 non-domain papers and their citations. However, two linguists are provided with all 40 papers for inter-annotator checking. In total, each paper is annotated by four annotators. A similar approach is adapted for corpus *CC1* as well. All six annotators are provided with a training session to use the developed annotation tool and the knowledge of *CCRO* classes and their meanings. After the complete annotation on corpora *CC1* and *CC2*, reconciliation is made to formulate a gold standard with an inter-annotator agreement of over 90%.

The annotation tool is a collaborative application, easily adaptable that provides a simple mechanism to annotate citation texts in both corpora to *CCRO* properties.

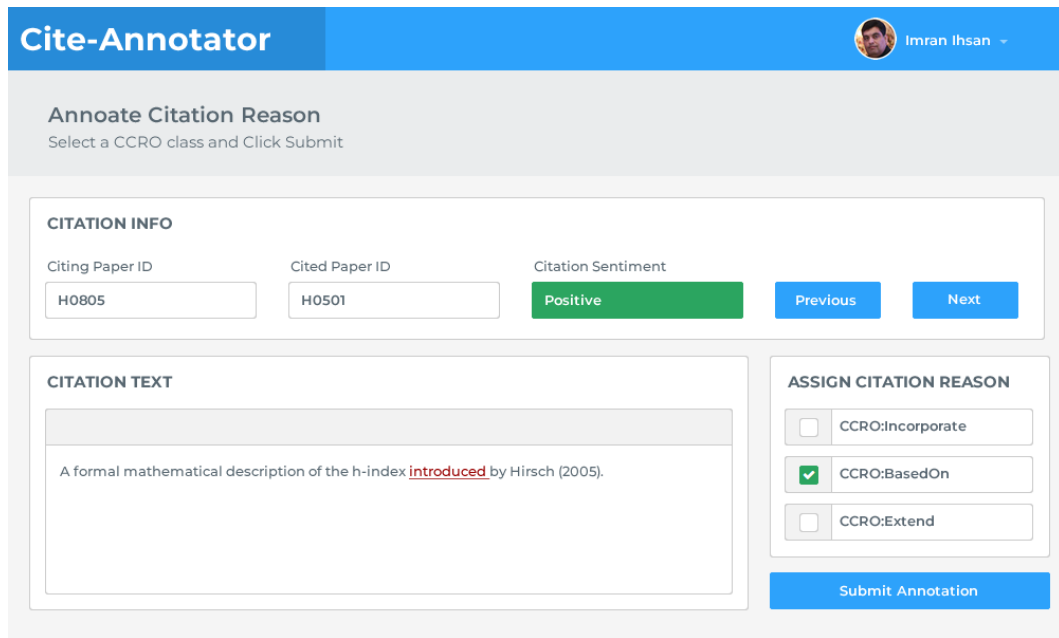


FIGURE 5.10: CCRO Annotation Tool Interface

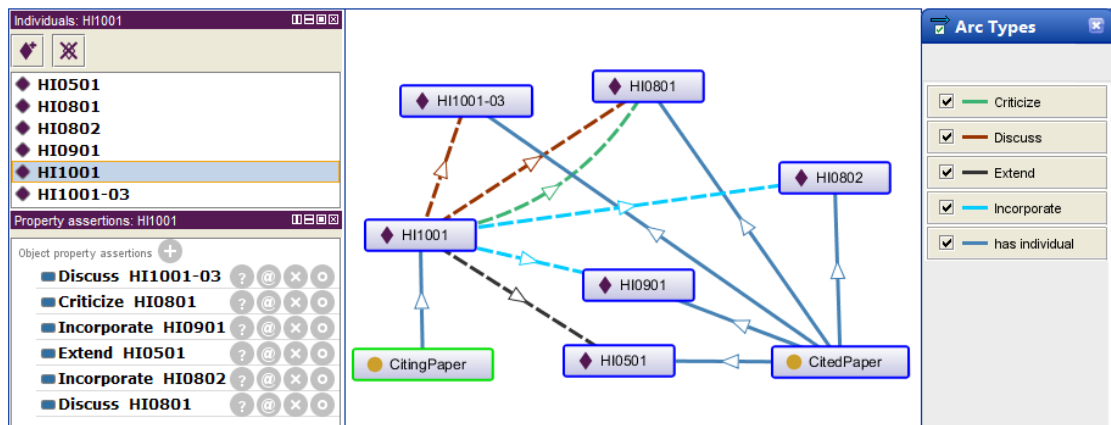


FIGURE 5.11: CCRO Instance for a Citing Paper - HI1001

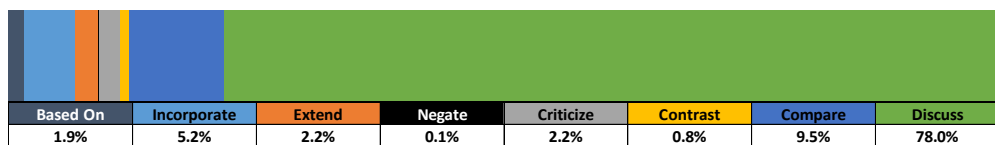


FIGURE 5.12: CCRO Properties Distribution (Corpus CC1 - 7390 Citations)

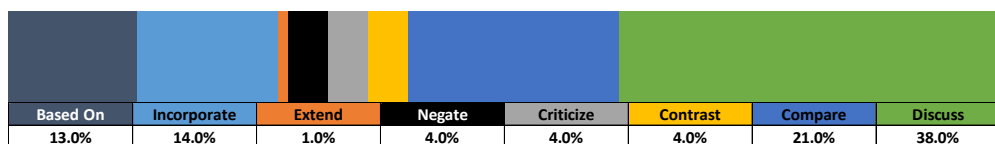


FIGURE 5.13: CCRO Properties Distribution (Corpus CC2 - 1230 Citations)

TABLE 5.3: CCRO Properties Distribution

CCRO Property	Corpus CC1	Corpus CC2
ccro:BasedOn	1.9%	13%
ccro:Incorporate	5.2%	14%
ccro:Extend	2.2%	1%
ccro:Negate	0.1%	4%
ccro:Criticize	2.2%	4%
ccro:Contrast	0.8%	4%
ccro:Compare	9.5%	21%
ccro:Discuss	78.0%	38%

Each annotator is provided with credentials to access the tool. After login, the tools display one citation at a time for annotation. Annotators can navigate between various assigned citations using “Next” and “Previous” buttons. The tool shows “Citing Paper ID”, “Cited Paper ID”, “Citation Sentiment”, and “Citation Text”. It also highlights *Reporting Verb* in a citation text. As both corpora have integrated citation’s sentiment polarity, the annotators are provided with a subset of *CCRO* classes corresponding to the sentiment polarity of the citation text rather than providing all 8 properties. If the sentiment of a citation text is “Positive”, provision of all eight citation reasons can cause annotators to annotate in “Negative” or “Neutral” sentiment, creating contradictions. Therefore, only the properties conforming to citation’s sentiment analysis are displayed. An author can choose, best-suited citation reason, and submits to move to the next annotation. Interface of annotation tool³, visible to annotators is shown in Figure 5.10, and Figure 5.11 represents one instance of a citing paper with ID ”HI1001” along with its cited papers.

³CCRO Annotation Tool: [https://github.com/imranihsan/CCRO/tree/master/ Annotation-Tool](https://github.com/imranihsan/CCRO/tree/master/Annotation-Tool)

After the complete annotation process, a distribution of *CCRO* Citation Reason Properties on Corpus *CC1* is shown in Fig 5.12. The results show 1.9% citation texts with “ccro:BasedOn”, 5.2% with “ccro:Incorporate”, 2.2% with “ccro:Extend”, 0.1% with “ccro:Negate”, 2.2% with “ccro:Criticize”, 0.8% with “ccro:Contrast”, 9.5% with “ccro:Compare” and 78% with “ccro:Discuss”. Fig 5.13 shows the *CCRO* Citation Reason Properties on Corpus *CC2*. Its results show 13% citation texts with “ccro:BasedOn”, 14% with “ccro:Incorporate”, 1% with “ccro:Extend”, 4% with “ccro:Negate”, 4% with “ccro:Criticize”, 4% with “ccro: Contrast”, 21% with “ccro:Compare” and 38% with “ccro:Discuss”. Table 5.3 also summarizes the results.

5.4 Results

To find the accuracy of the automatic distribution of *CCRO* properties on both corpora, the results are compared with the gold standard (manual annotation of citations’ reasons). A closer examination of automatic mapping shows that the algorithm has 53% accuracy in Positive sentiment, 72% in Negative, and 89% in Neutral for Corpus *CC1*. The algorithm has an overall accuracy of 85.4% with 0.74 Precision, 0.85 Recall, and 0.79 of F-Measure for *CC1*. The algorithm shows 70% accuracy in Positive sentiment, 76% in Negative, and 99% in Neutral for Corpus *CC2*. The algorithm shows better accuracy of 96.6% with 0.94 Precision, 0.96 Recall, and 0.95 of F-Measure for *CC2*. Fig 5.14 shows the tabulated results for both corpora against automatic mapping algorithms along with respective Precision, Recall and F-Measure. The results are calculated using four-cell contingency table [112, 113].

5.5 Findings

The success of the automatic mapping in each sentiment and *CCRO* property class has been tested against manual annotation gold standard.

Automatic Mapping of Citations' Reasons																	
CC1 7390 Citations	Based On	Incorporate	Extend	Negate	Criticize	Contrast	Compare	Discuss	CC2 1230 Citations	Based On	Incorporate	Extend	Negate	Criticize	Contrast	Compare	Discuss
	Based On	0	140	0	0	0	0	0		0	Based On	151	0	0	0	0	0
Incorporate	0	362	0	0	0	0	0	0	Incorporate	0	164	0	0	0	0	0	0
Extend	0	173	0	0	0	0	0	0	Extend	0	47	11	0	0	0	0	0
Negate	0	0	0	0	9	0	0	0	Negate	0	0	0	46	6	0	0	0
Criticize	0	0	0	0	169	0	0	0	Criticize	0	0	0	0	45	0	0	0
Contrast	0	0	0	0	58	0	0	0	Contrast	0	0	0	0	14	50	0	0
Compare	0	0	0	0	0	0	0	702	Compare	0	0	0	0	0	0	245	9
Discuss	0	0	0	0	0	0	0	5775	Discuss	0	0	0	0	0	0	0	442
Correct	Incorrect	Precision	Recall	F-Measure	Correct	Incorrect	Precision	Recall	F-Measure								
85.4%	14.6%	0.74	0.85	0.79	96.6%	3.4%	0.94	0.96	0.95								

FIGURE 5.14: Tabulated Results for Both Corpora CC1 and CC2

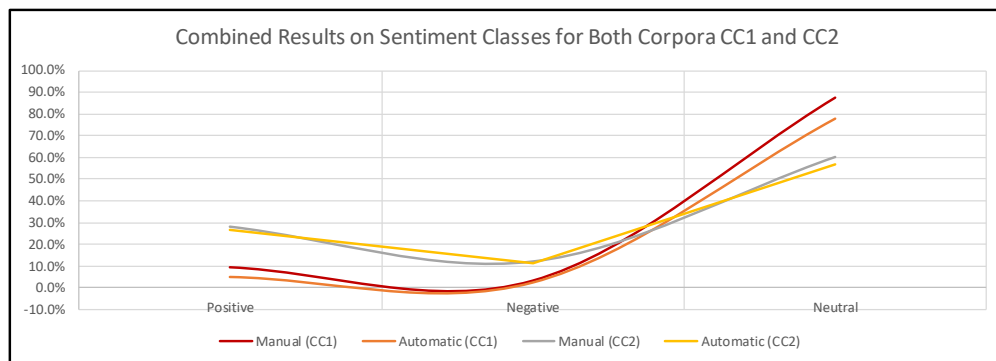


FIGURE 5.15: Combined Results on Sentiment for Both Corpora CC1 and CC2

5.5.1 Combined Results Based on Sentiment

Fig 5.15 represents a graphical representation of combined results for automatic mapping on all three on sentiments (Positive, Negative, and Neutral) for both corpora *CC1* and *CC2*. Results show that automatic mapping has performed better for corpus *CC2* as compared to *CC1* for “Positive” sentiment, however for “Negative” and “Neutral” sentiments, both corpora have shown similar behavior.

5.5.2 Combined Results Based on CCRO

Fig 5.16 represents a graphical representation of combined results for automatic mapping on *CCRO* citation reason properties for both corpora *CC1* and *CC2*.

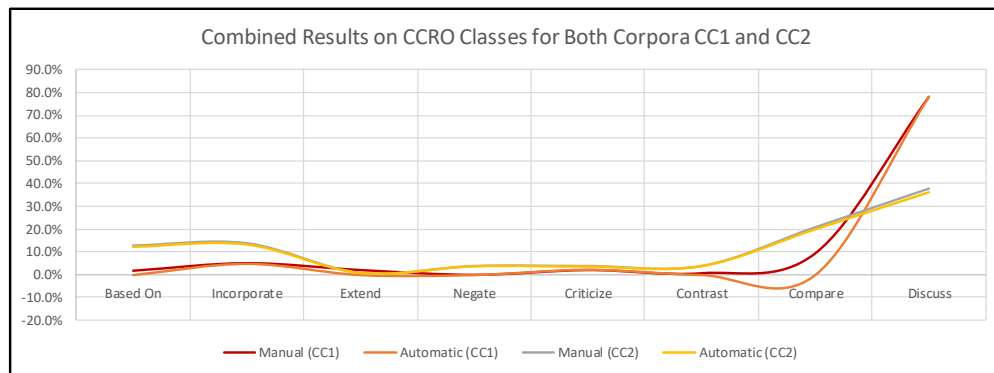


FIGURE 5.16: Combined Results on CCRO for Both Corpora CC1 and CC2

Results show that automatic mapping has shown poor results for “Based On”, “Extend”, “Negate”, “Contrast” and “Compare” for corpus CC1. However, for corpus CC2, the same algorithm has shown better results.

5.5.3 Limitations in Automatic Mapping

Automatic Mapping of CCRO properties initially require a citation to be classified using sentiment analysis. However, sentiment analysis of citation text is relatively difficult as compared to sentiment analysis of a product review [57]. One possible reason for this might be the writing style of a research paper where author restricts himself/herself to express personal opinions and avoids biased views. Additionally, a citation can have a wider range (from a single sentence to multiple paragraphs), making the sentiment analysis more difficult [57].

After sentiment analysis, automatic mapping of CCRO properties require investigation of linguistic patterns and rhetorical structures to find the citation’s reason. A study suggests [114] that there can be several limitations for automatic mapping of citation reasons, such as identification of citation function (*Reporting Verbs*) and establishment of one-to-one relationship between linguistic patterns and common-verb classes. However, according to Mercer and Di Marco [115] cue phrases (*Reporting Verbs*) exist in a citation context and can be extracted automatically. Therefore, there is a need to develop a stronger and concrete definition

of citation's function and inclusion of machine learning algorithms to achieve high precision and recall.

5.6 Generate Semantic Knowledge Graph

Linked Open Data has a pivotal role in semantic knowledge graphs. The knowledge graph can handle large data sets from various sources and link them to open data. This provides richer queries, knowledge discovery, and data analysis. Therefore, to convert both corpora, annotated with *CCRO* properties, an application is designed that reads each record from both corpora and create an *RDF* file using the guidelines provided in "*Citations' Context and Reasons Ontology - CCRO*". The transformation process consists of four (4) steps, that are;

1. Read each line from *CCRO* annotated Semantic Corpus
2. Parse and Extract Semantic Citations and *CCRO* Reasons
3. Fetch *IDs* for Citing and Cited Paper from Semantic Corpus
4. Generate *RDF* Files using schema defined in *CCRO*

After the complete process on both corpora, an *RDF Triple Store* is formed. The application is developed using *Microsoft .NET Framework*. To generate *RDF* Files, an open-source *.NET* library for *RDF*, known as "dotNetRDF⁴" is used. The library provides *APIs* for parsing, managing, querying, and writing *RDF* and *RDF Triple Stores*. A sample *RDF* file is shown in Figure 5.17.

5.7 Visualize Results

Based on the results, both corpora are converted into a *CSV* file format to be used in *R* as input for *iGraph*⁵ library. Citing paper and cited paper *IDs* from the

⁴dotNetRDF - <https://www.dotnetrdf.org/>

⁵<https://igraph.org/>

RDF FILE	<ccro:C04-1005>	a	ccro:CitingPaper	.
	<ccro:J93-2003>	a	ccro:CitedPaper	.
	<ccro:J97-3002>	a	ccro:CitedPaper	.
	<ccro:citation#1>	a	nif:string	;
		rdf:type	ccro:Citation	;
		biro:isReferencedBy	<ccro:C04-1005>	;
		biro:References	<ccro:J93-2003>	;
		nif:sentimentValue	-1	;
		ccro:MainVerb	"handle"	.
	<ccro:citation#2>	a	nif:string	;
		rdf:type	ccro:Citation	;
		biro:isReferencedBy	<ccro:C04-1005>	;
		biro:References	<ccro:J93-2003>	;
		nif:sentimentValue	0	;
		ccro:MainVerb	"introduce"	.
<ccro:citation#3>	a	nif:string	;	
	rdf:type	ccro:Citation	;	
	biro:isReferencedBy	<ccro:C04-1005>	;	
	biro:References	<ccro:J97-3002>	;	
	nif:sentimentValue	+1	;	
	ccro:MainVerb	"use"	.	
<ccro:C04-1005>	ccro:consistsOf	<ccro:citation#1>	;	
	ccro:Criticize	<ccro:J93-2003>	.	
<ccro:C04-1005>	ccro:consistsOf	<ccro:citation#2>	;	
	ccro:Discuss	<ccro:J93-2003>	.	
<ccro:C04-1005>	ccro:consistsOf	<ccro:citation#3>	;	
	ccro:Incorporate	<ccro:J97-3002>	.	

FIGURE 5.17: Sample RDF File

corpus becomes the nodes and the assigned CCRO class becomes the edge between these nodes. These edge classes are assigned different weights and colors in the adjacency matrix. *iGraph* library provides *walktrap.community*⁶ function to find densely connected sub-graphs. Using this idea, each *CCRO* class is identified as similar communities within the graph. The library provides a variety of layouts to visualize the graph, however, we have used ‘*The Davidson-Harel*’⁷ layout algorithm to visualize the graph. A partial representation of the visualized graph for corpus *CC1* is shown in Figure 5.18 and *CC2* (H-Index Domain) is shown in Figure 5.19.

⁶https://igraph.org/r/doc/cluster_walktrap.html

⁷https://igraph.org/r/doc/layout_with_dh.html

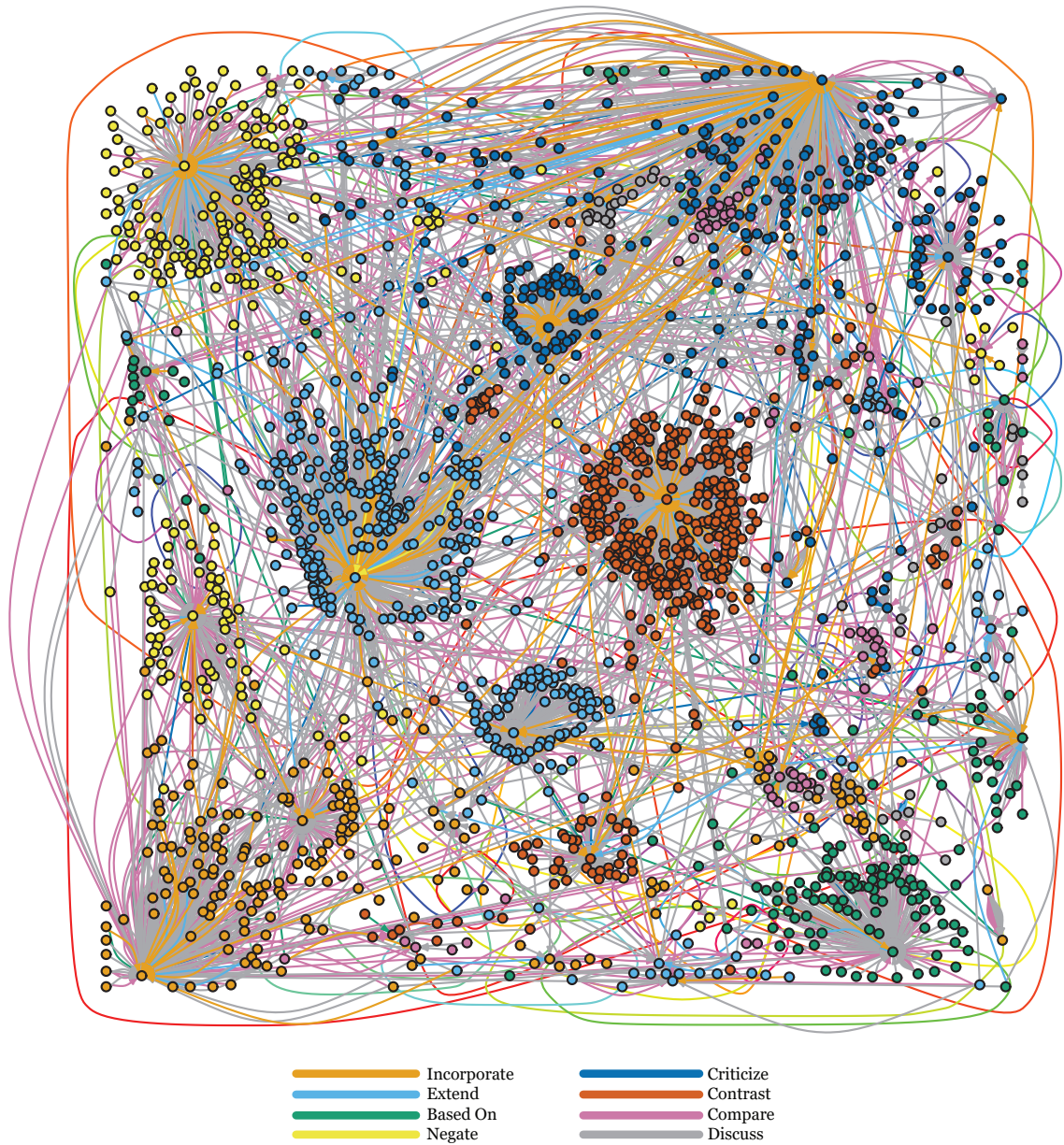


FIGURE 5.18: Ontology Based Semantic Citation Graph for Corpus CC1 - Partial Visualization

5.8 Conclusion

For instantiation and mapping of *Citation's Context and Reasons Ontology - CCRO* and its properties on real data, two approaches are used. First the automatic mapping of citation text using *Reporting Verb* and second human annotation with help of domain and linguist experts. Two types of data sets are employed in both experiments. One data set is a publicly available data set and is titled *Citation*

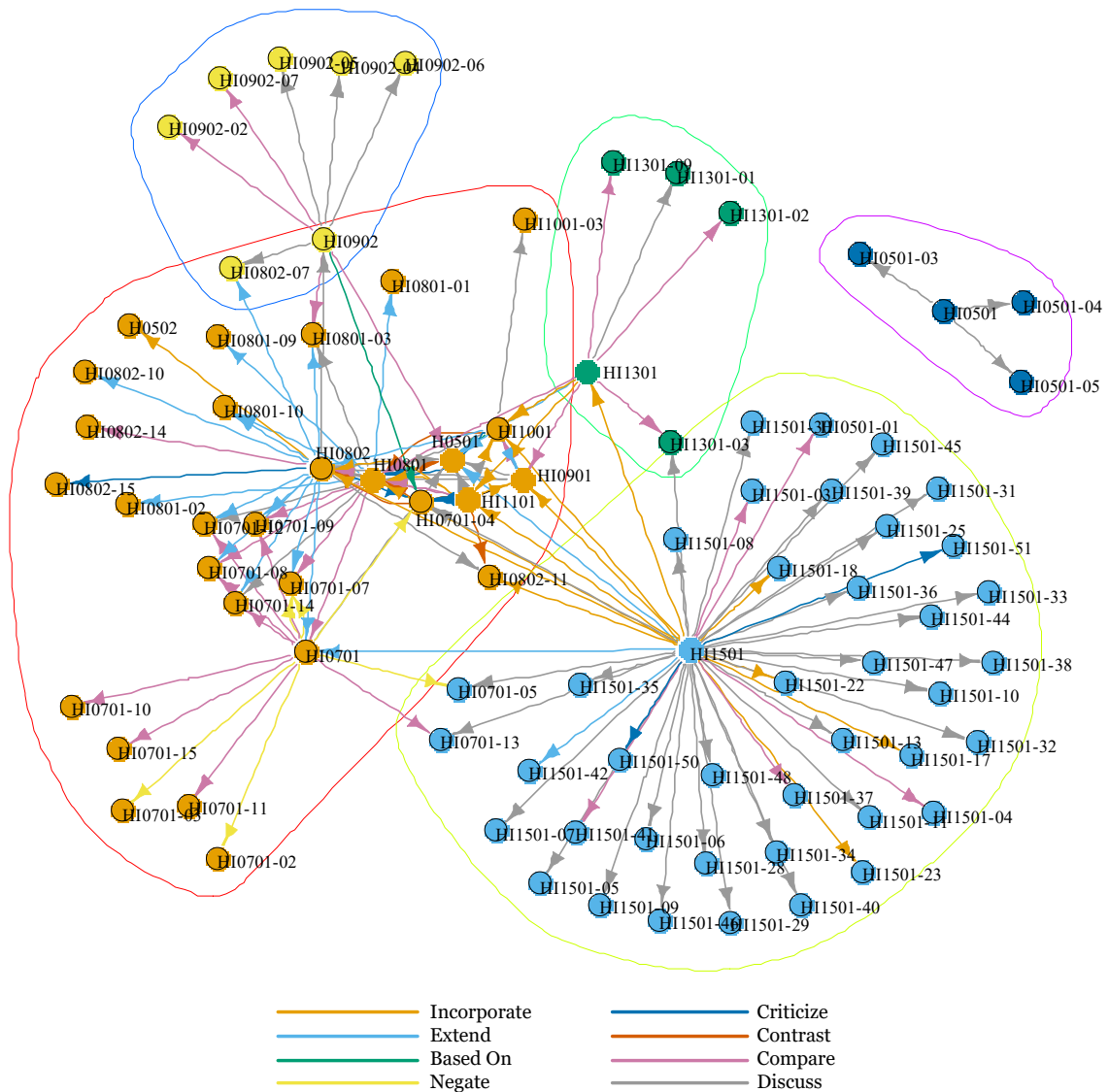


FIGURE 5.19: Ontology Based Semantic Citation Graph for Corpus CC2 - H-Index Domain

Corpus CC1 and the other is our own developed *Citation Corpus CC2* constituting citation texts from 40 influential research papers in four defined domains. Automatic mapping extracts *Reporting Verb* from a citation text and maps it onto a *CCRO* property using *Mapping Graph*, whereas, in human annotation, each expert is shown a citation text using an interface where they can annotate a citation text with a *CCRO* property. After processing both corpora, results are calculated and findings show that automatic mapping has shown 85.4% accuracy for CC1 and 96.6% for CC2. Afterwards, annotated corpora are converted into Linked Open Data using *RDF* and are visualized using *iGraph* in *R*.

There is a large volume of available scientific literature in each domain and authors spend a lot of time to search and filter the required information. Mapping of Citations' Context and Reasons Ontology can help in enriching the citation graph with meaningful tags and minimizing the time and effort to filter the required information. However, there is also a need to author a research article without destroying the useful semantics that can help automatic reasoners. Our survey (Section 2.4) on current research paper authoring and annotation tools and packages suggests that they do not provide integration of semantic and meaningful tags to define the context or reason of citation. Therefore, we have defined a system where the semantic model (ontology) of citations' reasons can be integrated within a futuristic research paper authoring tool that can help an author to choose semantic and meaningful tags for citation.

Chapter 6

CCRO Application: Query Semantic Graph

A scientific paper contains valuable information about the scholarly activity and its evolution. A citation graph has the potential to reveal important and interesting information about the history of particular scholarly research that has happened during its life-cycle. Citation graphs can be enriched with semantic tags, where scientific papers are inter-connected with citation reasons. Using semantic citation graphs, it is possible to infer the evolution of a research area over time, measure relations between research areas, and trace the influence of ideas that appear in the literature. Using *CCRO*'s citation reason class, known as “Based On”, it becomes possible to find how a main algorithm or concept has started or evolved.

There can be numerous applications that can be explored. However, we have discussed one application using the guidelines provided by Shum [32] to discover the evolutionary paths of scholarly activity, and answers to the following queries can provide a valid semantic model for such application. These queries are:

1. What is the current state of debate on this question?
2. Which theory author disagrees with?
3. What assumptions does this approach depend on?

4. Are there different schools of thought around this problem?
5. How many researches that have negated some other work ?

After the development of Linked Open Data for both corpora in the form of *RDF Triple Store*, *SPARQL* (A Query language for RDF Triple Stores) queries can be developed to look for the answers to the above questions. In the coming section, *SPARQL* query is developed for each question and the results are shown in a tabular form. Also, visualization is provided to show the results in semantic knowledge graph using the domain visualization techniques [64] for scientometrics. Automated research analysis using graph networks is now gaining popularity. Several pieces of research [116] are available that generalize convolution to graphs to conduct experiments on representation in large-scale graphs. However, the purpose of this visualization is to show a proof of concept for the developed ontology. Therefore, a simpler visualization technique is used. Let's investigate each query in detail.

6.1 Semantic Queries

6.1.1 What is the current state of debate on this question?

To test this query, an example is formulated on a well-known algorithm to measure both the productivity and citation impact of the publications of a scientist or scholar, known as "h-index". Two distinct papers are taken in the domain of Computer Science that represent the start and the current state of "h-index". These two papers and their assigned IDs are;

1. **H0501** – "An index to quantify an individual's scientific research output", (2005) by J. E. Hirsch [117]
2. **H0801** – "Completing h", (2015) by Keith R. Dienes [118]

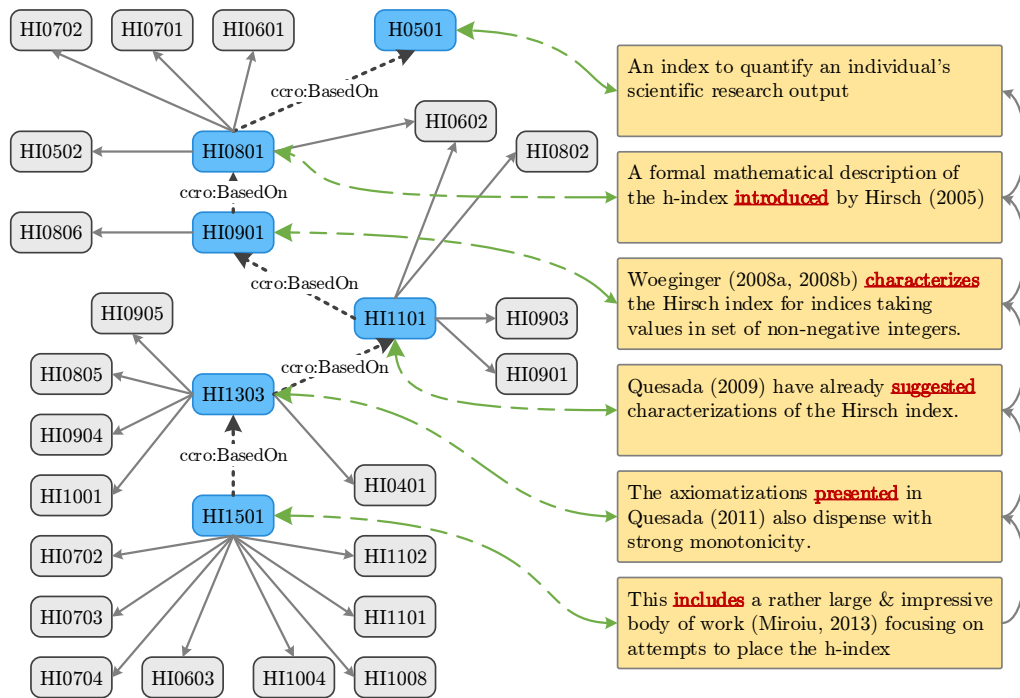


FIGURE 6.2: SPARQL Query 1: Results Visualization

The same experiment is then repeated on the Semantic Version of Citation Corpus *CC2*. To query *RDF Triple Store*, “Virtuoso Universal Server¹” is used. Virtuoso provides a middleware and database engine to load and query *RDF* data. Using the guidelines for the digital libraries, the *SPARQL* query is developed. This semantic query is:

```

PREFIX : <http://ccropus/resource/>
PREFIX ccro: <http://ccropus/ontology/>
SELECT ?paper
WHERE {
    ?p ccro:CitedPaper "H0501".
    ?paper ccro:Basedon* ?p.
}

```

Results of *SPARQL* Query 1 are shown in Table 6.1 with its visualization in Fig 6.2 along with their citation texts in the selected path, for a deeper understanding

¹Virtuoso - <https://virtuoso.openlinksw.com/>

TABLE 6.1: SPARQL Query 1: Results

?Paper	Paper Title for Information Only
HI0501	//An index to quantify an individual’s scientific research output
HI0801	//An axiomatic characterization of the Hirsch-index
HI0901	//Monotonicity and the Hirsch index
HI1101	//Further characterizations of the Hirsch index
HI1301	//Axiomatizing the Hirsch index: Quantity and quality disjointed
HI1501	//Completing h

and to help the author to see the evolution of an algorithm or research. From both Fig 6.1 and Fig 6.2, it is evident that both experiments resulted in a similar scholarly path between the selected articles.

6.1.2 Which theory author disagrees with?

For *SPARQL* Query 2, the research paper chosen is “HI0701”, titled “Does the h-index have predictive power?”. In this research paper, the author has argued about the drawbacks of some “H-Index” modifications. After applying the query on Semantic version of Citation Corpus *SCC2* using the *CCRO* property “Negate”, the results are shown in Table 6.2 and visually illustrated in Figure 6.3.

```

PREFIX : <http://ccropus/resource/>
PREFIX ccro: <http://ccropus/ontology/>
SELECT ?paper
WHERE {
    ?p ccro:CitingPaper "HI0701";
    ccro:Negate ?paper.
}

```

TABLE 6.2: SPARQL Query 2: Results

?Paper	Paper Title for Information Only
HI0701-02	//A quantitative analysis of measures of quality in science
HI0701-03	// Measures for measures
HI0701-04	//Monotonicity and the Hirsch index
HI0701-05	//Theory and practise of the g-index
HI0701-07	//What do we know about the h index?

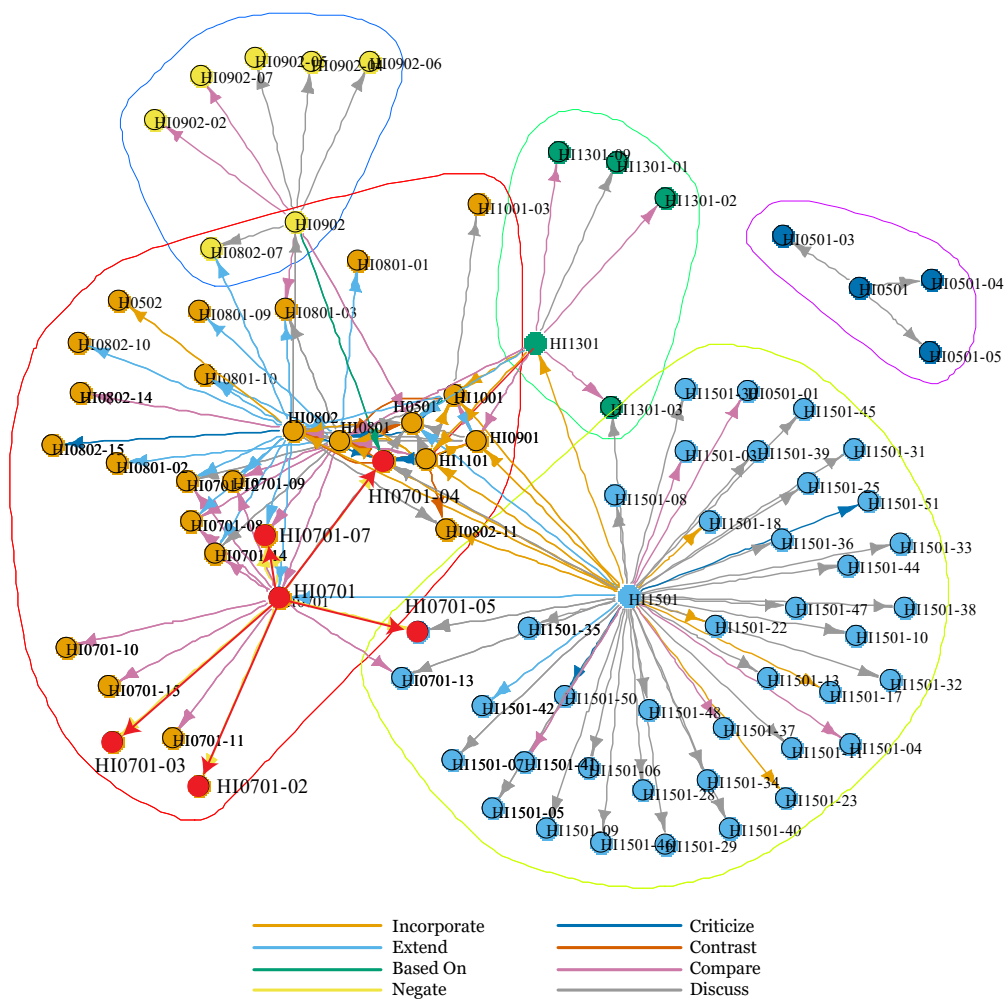


FIGURE 6.3: Citation Sub Graph for Query 2 using Citation Corpus SCC2

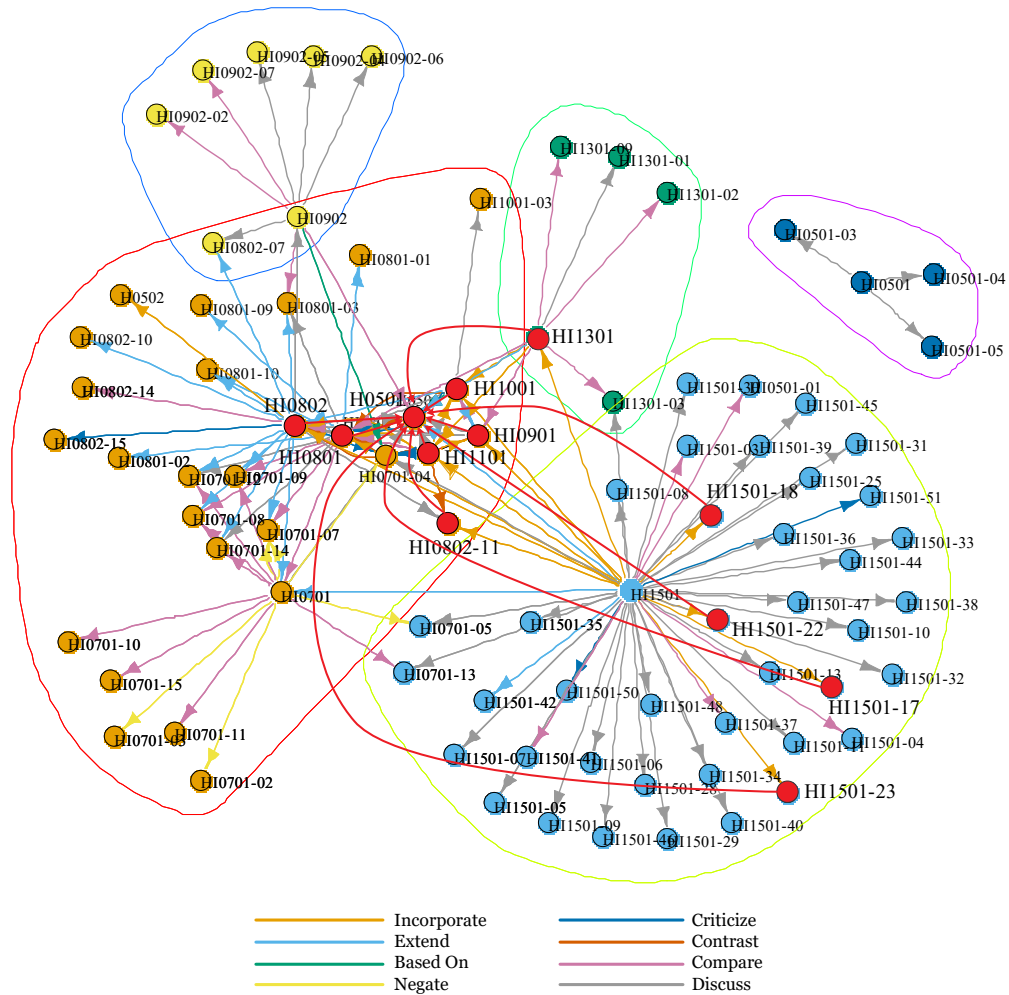


FIGURE 6.4: Citation Sub Graph for Query 3 using Citation Corpus SCC2

6.1.3 What assumptions does this approach depends on?

Research papers tend to cite some important theories that are necessary to build the new research forming the basis of assumptions that are required for the new approach. Using *CCRO* property “Based on” for the citation in the selected paper, a query has been formulated to find out how many research papers in our selected corpus have used “H0501 - Completing h” as the basis of their approach. After formulating and applying the query on the Semantic version of Citation Corpus *SCC2*, the results are shown in Table 6.3 and illustrated in Figure 6.4.

```

PREFIX : <http://ccropus/resource/>
PREFIX ccro: <http://ccropus/ontology/>
SELECT ?paper

```

TABLE 6.3: SPARQL Query 3: Results

?Paper	Paper Title for Information Only
HI0801	//An axiomatic characterization of the Hirsch-index
HI0802	//A symmetry axiom for scientific impact indices
HI0802-11	//An axiomatic characterization of the ranking based on the h-index...
HI1501-17	//Consistent bibliometric rankings of authors and of journals
HI1501-18	//An axiomatic approach to bibliometric rankings and indices
HI0901	//Monotonicity and the Hirsch index
HI1001	//More axiomatics for the Hirsch index
HI1101	//Further characterizations of the Hirsch index
HI1501-22	//Axiomatics for the Hirsch index and the Egghe index
HI1501-23	//An axiomatization of the Hirsch-index without adopting monotonicity
HI1301	//Axiomatizing the Hirsch index: Quantity and quality disjointed

```

WHERE {
    ?p ccro:CitedPaper "H1501".
    ?paper ccro:Basedon ?p.
}

```

6.1.4 Are there different school of thoughts around this problem?

A different school of thought means that research either agrees or disagrees with existing research. Using the research paper “H0501 - Completing h”, *SPARQL*

TABLE 6.4: SPARQL Query 4: Results

?Positive	?Negative	Paper Title for Information Only
HI1001		//More axiomatics for the Hirsch index
HI1301		//Axiomatizing the Hirsch index
HI1501		//Completing h
HI0802		//A symmetry axiom for scientific impact indices
	HI0801	//An axiomatic characterization of the Hirsch-index

Query 4 is formulated to find research papers that agree with the approach defined to access the quality of a research paper or disagree with the approach altogether. Using the properties in *CCRO* for “Positive” and “Negative” sentiment, two different schools of thought can be extracted with “Positive” showing papers that agree with the approach while “Negative” showing papers that are against it. After applying the query on the Semantic version of Citation Corpus *CC2*, the results are shown in Table 6.4 and illustrated in Figure 6.5.

```

PREFIX : <http://ccropus/resource/>
PREFIX ccro: <http://ccropus/ontology/>
SELECT ?positive ?negative
WHERE {
    ?positive ccro:Incorporate "H1501";
              ccro:Extend "H1501";
              ccro:Basedon "H1501".
    ?Negative ccro:Negate "H1501";
              ccro:Criticize "H1501";
              ccro:Contrast "H1501". }

```

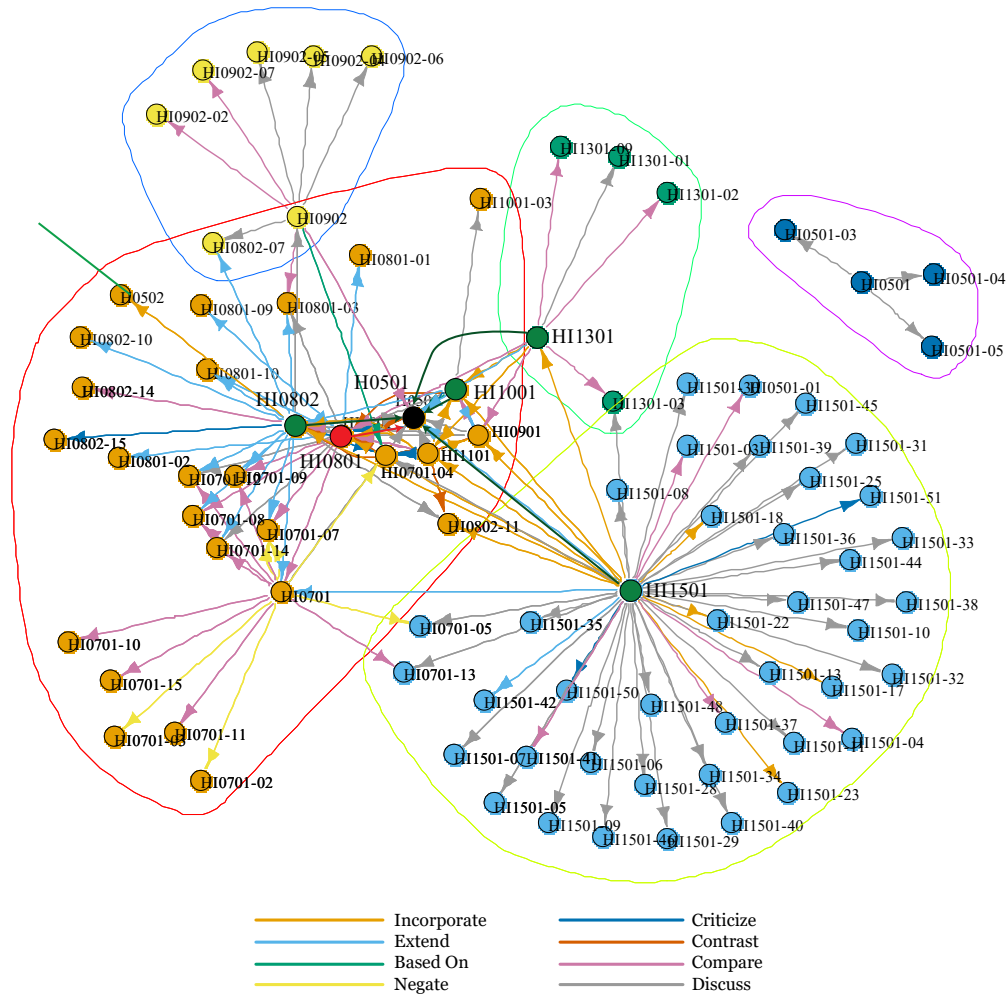


FIGURE 6.5: Citation Sub Graph for Query 4 using Citation Corpus SCC2

6.1.5 How many researches that have negated some other work?

SPARQL Query 5 finds the number of research papers that have negated any other research. Using the *CCRO* property “Negate” on the Semantic version of Citation Corpus *CC2*, all the citations using this property are extracted. After applying an aggregate function “Count”, the results are tabulated and the results are shown in Table 6.5.

```
PREFIX : <http://ccropus/resource/>
PREFIX ccro: <http://ccropus/ontology/>
SELECT COUNT(?paper)
```

TABLE 6.5: SPARQL Query 5: Results

?Count
9

```
WHERE {  
    ?paper ccro:Negate ?p.  
}
```

6.2 Conclusion

For a proof of concept for *Citation's Context and Reasons Ontology - CCRO*, five queries are formulated. These queries are implemented on the *RDF* based Semantic Version of Citation Corpus *CC2*. The results are tabulated and visualized. Though the experiment shows promising results, there is a need to find the degree of generalization in large-scale corpora of various disciplines. As a future direction, we are working on a crawler algorithm that can extract citation texts from various research papers on a large scale and its conversion to a semantic graph using CCRO.

Chapter 7

CCRO Application: Semantic Authoring

The reason to write any scientific scholarly document is to advance the accumulated knowledge in a verifiable way. Authors communicate this knowledge through literature review to form and present scientific claims along with their justifications. The most common method adopted by the authors to form the discourse and express in a document is via citation. We have discussed the rhetorical and argumentative nature of such discourse by many researchers (Chapter 2) in past by providing insights into why authors cite specific research and a need for a semantic-based research paper authoring tool that can help an author to choose semantic and meaningful tags for citation.

Semantic-based authoring can create an ecosystem to alleviate the information overload problem. The scientific community either use \LaTeX or traditional processing systems for authoring research papers. Traditional processing systems include Microsoft Word, Google Docs, LibreOffice, Apple Pages, etc. Our solution relies on enriching scientific publications with explicit rhetorical and argumentation discourse structures, using Citation's Context and Reasons Ontology - *CCRO* by identifying and classifying citation texts within \LaTeX files. So, the question is, how many authors use \LaTeX typesetting language for authoring. To answer

TABLE 7.1: Summary Statistics of use of L^AT_EX in Science Disciplines

Discipline	L ^A T _E X Rate
Mathematics	96.9
Statistics and Probability	89.1
Physics	74.0
Computer Science	45.8
Astronomy and Astrophysics	35.1
Engineering	1.0
Geosciences	0.8
Ecology	0.4
Chemistry	0.3
Biology	0.0
Medicine	0.0
Psychology	0.0
Sport Sciences	0.0
Average/Mean	26.8

this, a study [94] was conducted to investigate the presentation and adoption of L^AT_EX across various disciplines. The data was extracted with the help of a paper called “Don’t Format Manuscripts” [119] and calculated using a compare tool Scopus/SciMago¹. Their results are shown in Table 7.1.

Using Scopus/SciMago, the total number of citable documents from 1996 to 2019 in the field of Mathematics, Physics, and Astronomy, and Computer Science, L^AT_EX articles are calculated. Table 7.2 shows the tabulated results.

¹SciMago: <https://www.scimagojr.com/countryrank.php>

TABLE 7.2: Summary Statistics of use of L^AT_EX in Science Disciplines

Discipline	Cited Documents	L ^A T _E X Rate	L ^A T _E X Documents
Mathematics	4,190,427	96.9	4,022,810
Physics and Astronomy	8,262,894	60.0	4,957,736
Computer Science	6,556,510	45.8	2,950,430
Total	19,009,831		11,930,976

Based on the study, around 27% researchers used L^AT_EX typesetting for authoring and an astonishing 11,930,976 citable documents are written using L^AT_EX in hard sciences. Another plus point for L^AT_EX typesetting is its availability as open-source as compared to the traditional processing system. Therefore, L^AT_EX typesetting is selected for the development of Semantic Publishing Ecosystem using explicit rhetorical and argumentation discourse structures. Furthermore, embedding these structures within *RDF* Data Store enables the creation of semantic publications that lay a foundation artifact for the Semantic Publishing Ecosystem and linked resources to become part of the current Web of Data.

7.1 Methodology

The complete process for Semantic Publishing Ecosystem spans over four steps, “Semantic Annotation”, “Semantic Authoring”, “Semantic Publishing”, and “Semantic Graph”. Fig 7.1 describes the methodology adopted.

7.1.1 Semantic Annotation

Based on our survey, it is evident to develop a citation package that can provide semantic annotation. Advance feature of L^AT_EX allows creation of *.ins* and *.dtx* files for creating and distributing classes and style files [120]. Using the *Natbib*

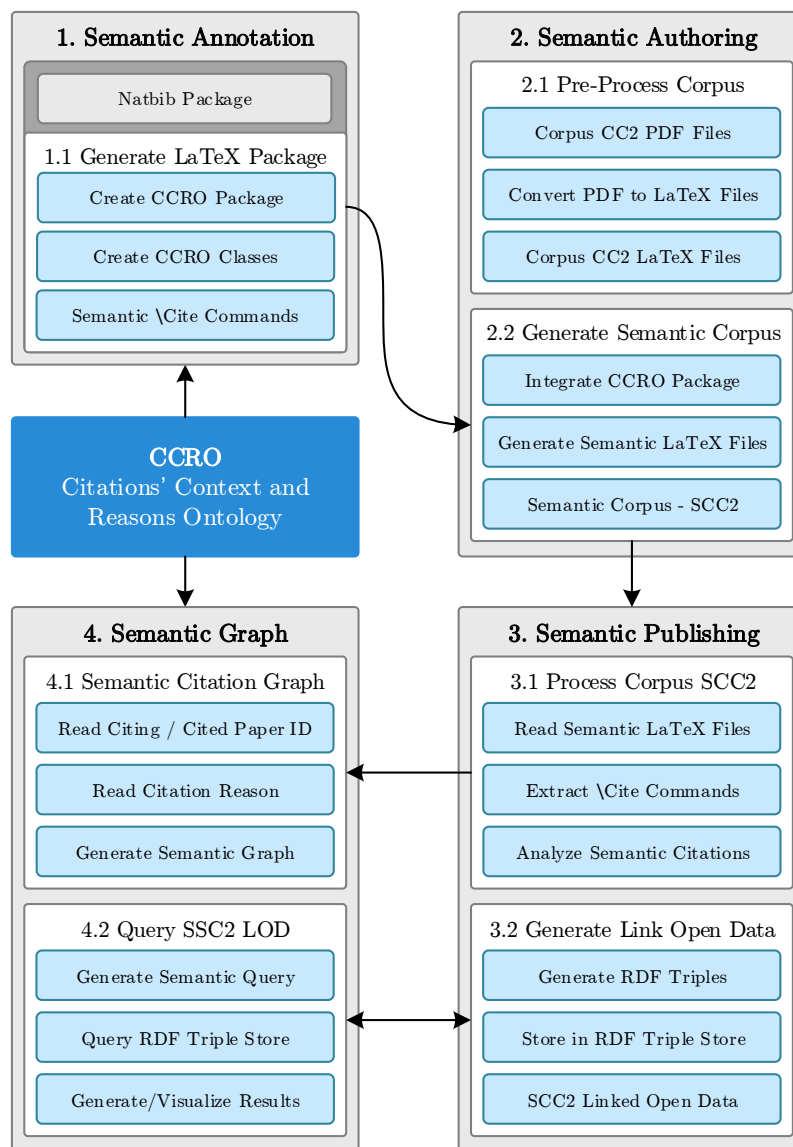


FIGURE 7.1: Steps towards Semantic Authoring and Publishing

Package, the *CCRO* Package is designed integrating Citation's Context and Reason's Ontology classes. Semantic Citation commands are created that can be used in any available \LaTeX Editor.

7.1.2 Semantic Authoring

Authors typically use \LaTeX to write their research articles. However, to the best of our knowledge, no repository publicly exists that houses \LaTeX files. In order

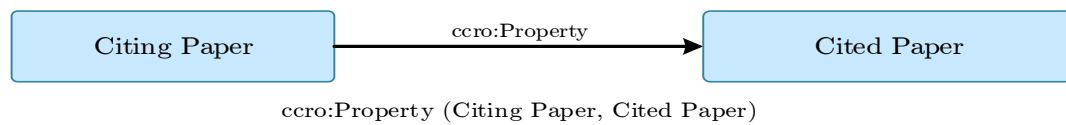


FIGURE 7.2: RDF Triple in RDF Data Store

to simulate the authoring environment, all 40 research papers from the Corpus *CC2* (Section 3.2.1.2) are selected and manually converted into \LaTeX files using standard \LaTeX template. The resultant is a collection of 40 \LaTeX files that act as our input corpus.

Using the developed *CCRO* Package for semantic citation, all 40 \LaTeX files in the selected corpus are manually converted into semantic-based \LaTeX files. Kindly note, each \LaTeX file is separately stored after the inclusion of semantic citation tags. In principle, the package enables authors to integrate semantic citation while authoring the paper and providing the reason why he/she is citing a paper. However, the basic concept of integrating a bibliographic entry in a \LaTeX file adapts the same procedure as described in the world known *Natbib* package.

7.1.3 Semantic Publishing

After converting the selected \LaTeX files into semantic \LaTeX files, the collection can be referred to as “Semantic Corpus - SCC2”. For Semantic Publishing, an application is developed that reads each semantic \LaTeX file from the Semantic Corpus, automatically extracts semantic citations, and converts them into an *RDF* Triple. As *RDF* Triple is composed of “Subject – Predicate – Object”, therefore each triple contains the Citing Paper as the “Subject”, the Cite Paper as the “Object” and the selected *CCRO* Property as the “Predicate” as shown in Figure 7.2. This collection of all *RDF* Triples is known as *RDF Triple Store* that formulates *Semantic Corpus SCC2 Linked Open Data*.

7.1.4 Semantic Graph

Semantic Graph can store information in a rich, contextual, and conceptual construct. This construct is commonly called a ‘triple’. Using the triples available in *Semantic Corpus SSC2 Linked Open Data* semantic graph is then visualized. Generated *RDF Triple Store* is thus a semantic graph that may contain valuable information regarding how a scholarly activity has evolved during its lifecycle. To find the evolutionary paths between the scholarly activity, *SPARQL* queries are written and executed on *RDF Data Store*. The results are then visualized for discourse analysis.

7.2 Experiment and Results

To create an ecosystem of semantic authoring and publishing, four experiments are performed. First experiment is to create semantic annotation by creating the *CCRO* Package for \LaTeX . The second is to integrate these semantic tags in \LaTeX files. The third is an application to automatically read semantic tags from \LaTeX files and create *RDF* Data Store and the Last experiment is to visualize *RDF* Data Store (Semantic Citation Graph) among selected papers.

7.2.1 The CCRO Package

The “*CCRO*” Package is an extension to \LaTeX “ $\backslash\text{cite}\{\text{paperID}\}$ ” command by integrating semantic-based citations. The package is based on “*Natbib*” Package and is compatible with the standard bibliographic style files such as “*harvard*”, “*apacite*” and “*chicago*” etc.

In contrast to other packages, the “*CCRO*” Package supports semantic tagging of citations. The Package uses Citation’s Context and Reasons Ontology - *CCRO*’s constituent properties to create a meaningful tag between the citing and the cited

paper. Like all other packages, it is required to be loaded in the document preamble such as

```
\usepackage{CCRO}
```

The document text itself begins with:

```
\begin{document}
\bibliographystyle{plainnat}
```

“plainnat” specifies the bibliography style used by the “BIBTEX” program to generate the actual bibliography from a database. The style “plainnat” is adapted from “natbib”. However, any other bibliographic styles can be used instead of “plainnat”

To make a semantic citation in the text, the following commands are formed

<code>\citepos{paperID}{+1}</code>	for semantic citation “ <i>ccro:Incorporate</i> ”.
<code>\citepos{paperID}{+2}</code>	for semantic citation “ <i>ccro:Extend</i> ”.
<code>\citepos{paperID}{+3}</code>	for semantic citation “ <i>ccro:BasedOn</i> ”.
<code>\citeneg{paperID}{-1}</code>	for semantic citation “ <i>ccro:Contrast</i> ”.
<code>\citeneg{paperID}{-2}</code>	for semantic citation “ <i>ccro:Criticize</i> ”.
<code>\citeneg{paperID}{-3}</code>	for semantic citation “ <i>ccro:Negate</i> ”.
<code>\citeneu{paperID}{=1}</code>	for semantic citation “ <i>ccro:Discuss</i> ”.
<code>\citeneu{paperID}{=2}</code>	for semantic citation “ <i>ccro:Compare</i> ”.

Where `\citepos` command defines citations’ reason classes in “Positive” context, `\citeneg` in “Negative” and `\citeneu` in “Neutral”. Fig 7.3 defines the syntax and its output in detail. Though, using numbers as commands is not user-friendly, including complete names such as `\citepos{PaperID}{Incorporate}` becomes laborious for authors. For the future version of *CCRO* package, we are working on a smarter way to incorporate citation reasons. The current package is provided using `ccro.sty` file and is available online at <https://github.com/imranihsan/CCRO/blob/master/ccro.sty> (Appendix: E)

<code>\citepos{Ihsan2019}{+1}</code>	[Incorporate Ihsan et al., 2019]
<code>\citepos{Ihsan2019}{+2}</code>	[Extend Ihsan et al., 2019]
<code>\citepos{Ihsan2019}{+3}</code>	[Basedon Ihsan et al., 2019]
<code>\citeneg{Ihsan2019}{-1}</code>	[Contrast Ihsan et al., 2019]
<code>\citeneg{Ihsan2019}{-2}</code>	[Criticize Ihsan et al., 2019]
<code>\citeneg{Ihsan2019}{-3}</code>	[Negate Ihsan et al., 2019]
<code>\citeneu{Ihsan2019}{=1}</code>	[Discuss Ihsan et al., 2019]
<code>\citeneu{Ihsan2019}{=2}</code>	[Compare Ihsan et al., 2019]

FIGURE 7.3: CCRO Package for L^AT_EX - Syntax and Output

7.2.2 Semantic L^AT_EX

Semantic L^AT_EX is the extension of L^AT_EX writing environment that supports the semantic annotation of citations based on citations' context and reasons using the CCRO Package. Semantic L^AT_EX lets the author choose a reason from the available list while citing another research in the content of the research paper. The process is more robust than as defined by SALT [33]. In Semantically Annotated L^AT_EX - SALT, semantic annotation is provided as metadata in the RDF file along with PDF document using *Annotation Ontology*. However, for Semantic L^AT_EX, the author does not create a separate RDF file for metadata, rather the file can be automatically created along with the process to generate PDF. A sample of Semantic L^AT_EX using CCRO Package is shown in Fig 7.4. Using this technique all 40 research papers are initially converted into L^AT_EX before extending them into Semantic L^AT_EX.

7.2.3 Semantic Publishing Process

The semantic publishing is an application that takes Semantic L^AT_EX documents as an input and creates a *RDF* file using the guidelines provided in *Citations' Context and Reasons Ontology - CCRO*. The transformation process consists of six steps, that are;

<code>\citeneq{J93-2003}{-2}</code>	For example, the statistical word alignment in IBM translation models [Criticize Brown et al., 1993] can only handle word to word and multi-word to word alignments.
<code>\citeneu{J93-2003}{=1}</code>	Bilingual word alignment is first introduced as an intermediate result in statistical machine translation (SMT) [Discuss Brown et al., 1993].
<code>\citepos{J97-3002}{+1}</code>	In addition, Wu [Incorporate Wu., 1997] used a stochastic inversion transduction grammar to simultaneously parse the sentence pairs to get the word or phrase alignments.

FIGURE 7.4: A Semantic L^AT_EX Sample

1. Read each **Semantic** L^AT_EX document from Semantic Corpus
2. Parse and Extract Semantic Citations
3. Analyze the Semantic Citations `\cite` commands
4. Fetch IDs for Citing and Cited Paper from Semantic Corpus
5. Generate RDF Files using schema defined in CCRO
6. Generate PDF Files using **CCRO** Package

Figure 7.5 shows a sample **Semantic** L^AT_EX document with its *Semantic RDF* and *PDF* counter parts. After the complete process on entire semantic corpus, an *RDF Triple Store* is generated. The application is developed using *Microsoft .NET Framework*. To generate *RDF* Files, an open source *.NET* library for *RDF*, known as “dotNetRDF²” is used. The library provides *APIs* for parsing, managing, querying and writing *RDF* and *RDF Triple Stores*.

7.2.4 Semantic Citation Graph

RDF Triple Store for **Semantic** L^AT_EX documents can be visualized as a semantic citation graph, describing a cognitive link between citing and cited paper. The

²dotNetRDF - <https://www.dotnetrdf.org/>

Semantic PDF	<p>For example, the statistical word alignment in IBM translation models [Criticize Brown et al., 1993] can only handle word to word and multi-word to word alignments.</p> <p>Bilingual word alignment is first introduced as an intermediate result in statistical machine translation (SMT) [Discuss Brown et al., 1993].</p> <p>In addition, Wu [Incorporate Wu., 1997] used a stochastic inversion transduction grammar to simultaneously parse the sentence pairs to get the word or phrase alignments.</p>
Semantic LATEX	<p>For example, the statistical word alignment in IBM translation models <code>\citeneq{J93-2003}{-2}</code> can only handle word to word and multi-word to word alignments.</p> <p>Bilingual word alignment is first introduced as an intermediate result in statistical machine translation (SMT) <code>\citeneu{J93-2003}{=1}</code>.</p> <p>In addition, <code>\citepos{J97-3002}{+1}</code>. used a stochastic inversion transduction grammar to simultaneously parse the sentence pairs to get the word or phrase alignments.</p>
Semantic RDF	<pre> <ccro:C04-1005> a ccro:CitingPaper . <ccro:J93-2003> a ccro:CitedPaper . <ccro:J97-3002> a ccro:CitedPaper . <ccro:citation#1> a nif:string ; rdf:type ccro:Citation ; biro:isReferencedBy <ccro:C04-1005> ; biro:References <ccro:J93-2003> ; nif:sentimentValue -1 ; ccro:MainVerb "handle" . <ccro:citation#2> a nif:string ; rdf:type ccro:Citation ; biro:isReferencedBy <ccro:C04-1005> ; biro:References <ccro:J93-2003> ; nif:sentimentValue 0 ; ccro:MainVerb "introduce" . <ccro:citation#3> a nif:string ; rdf:type ccro:Citation ; biro:isReferencedBy <ccro:C04-1005> ; biro:References <ccro:J97-3002> ; nif:sentimentValue +1 ; ccro:MainVerb "use" . <ccro:C04-1005> ccro:consistsOf <ccro:citation#1> ; ccro:Criticize <ccro:J93-2003> . <ccro:C04-1005> ccro:consistsOf <ccro:citation#2> ; ccro:Discuss <ccro:J93-2003> . <ccro:C04-1005> ccro:consistsOf <ccro:citation#3> ; ccro:Incorporate <ccro:J97-3002> . </pre>

FIGURE 7.5: A Semantic Publishing Process

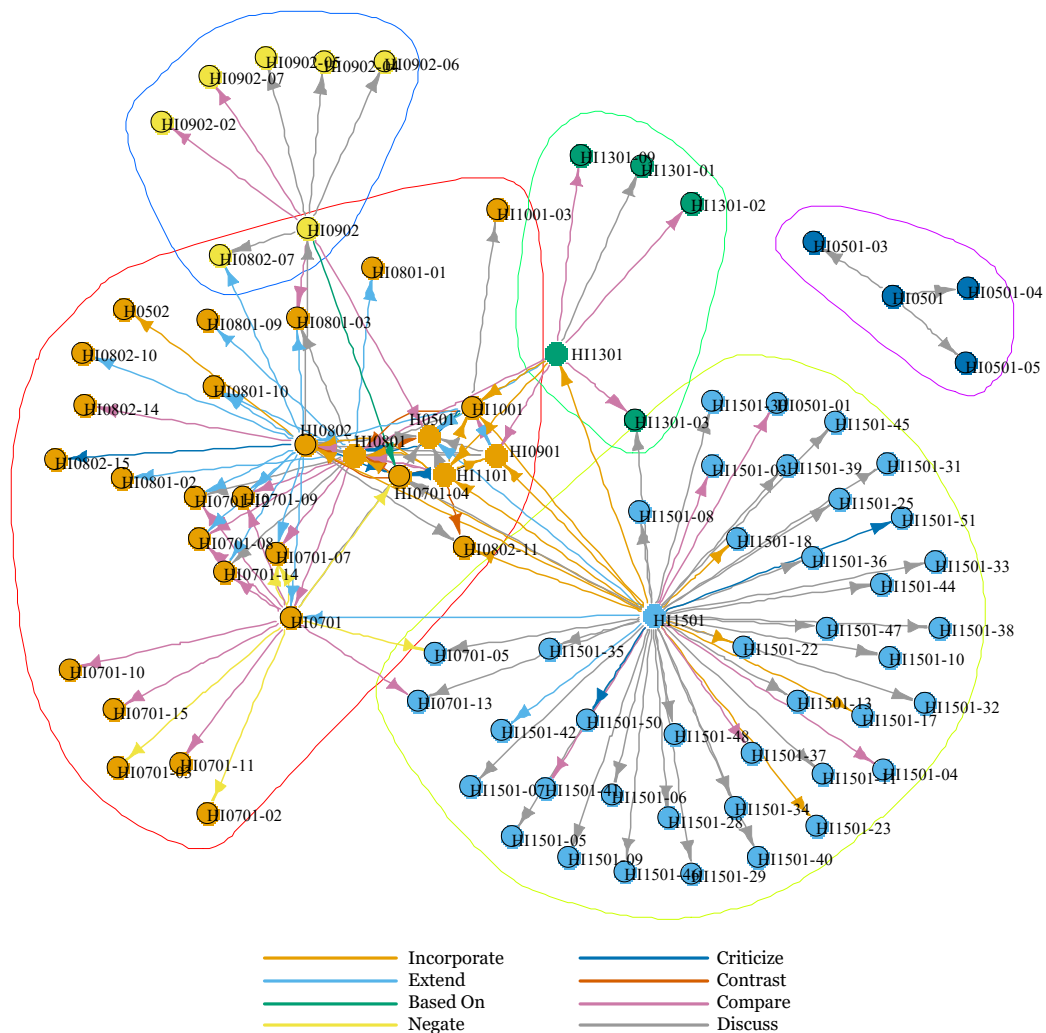


FIGURE 7.6: CCRO Based Semantic Citation Graph

visualization uses the citing paper and cited paper IDs from semantic corpus as nodes and the assigned CCRO class as the edge between these nodes. These edge classes are assigned different weights and colors. A partial representation of the visualized graph for the semantic corpus is shown in Figure 7.6.

7.3 Conclusion

One of the best sources of knowledge to tell the reason for citation is the author of a paper at the time when he/she is writing the paper. Authors of the scholarly articles cite other articles based on certain reasons. We have developed a semantic

annotation package for L^AT_EX to integrate citation reasons at the time of authoring a paper and this package can be integrated into any L^AT_EX authoring tool. With the help of this package, authors can tag a citation using suitable *CCRO* properties, making a simple L^AT_EX document as **Semantic L^AT_EX** document. Afterward, semantic citation tags embedded in a **Semantic L^AT_EX** document can be stored in a *RDF Triple Store* to formulate a semantically enriched citation graph using citations' context and reasons where Berners-Lee's [121] vision of semantic web and giving meanings to hyperlinks can be adapted in its true essence for scholarly publishing.

Development of semantically enriched and machine-understandable citation graphs can become the foundation for many applications, such as the discovery of evolutionary paths in scholarly activity or finding influential papers within a certain domain. In the next chapter, we have investigated one of the many possible applications of semantic authoring and publishing.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

Keeping in mind our research questions (Section 1.8), surveys, experiments and results, we conclude the thesis with following list of major conclusions.

1. A scientific research paper contains vital information that incites its citation by the authors and researchers, based on diverse reasons. These reasons become important parameters to discover cognitive relations between research papers. Automated processing of the citations' data requires a formal and semantic definition of the citation reasons. The study has revealed that there have been many attempts to record citations in a semantically meaningful way in the form of ontology. However, most of these ontologies do not deal with citation reasons. One ontology (*Citation Type Ontology - CiTo*) provides a formal and semantic definition of citation reasons but it has limitations. These limitations include “Less and Most Used Neutral Properties”, “Lower Inter-Rate Agreement”, “Non-Taxonomic Organization of Properties”, “Misinterpretation of Properties” and “Properties Perspective”. In general, CiTO depicts the perspective of an annotator and not the author.

2. We have discovered and evaluated more than 150 citations' reasons from the existing published literature to be represented as citation tags. Many of these reasons have overlapped and diffused meanings. Annotating such a large volume of citation graphs with citation's reasons manually is nearly impossible. Thus, by adopting a well-defined scientific methodology (methonotology) to formulate an ontology of citation reasons, we have reduced 150 reasons into only eight, using an iterative process of sentiment analysis, collaborative meanings, and experts' opinions.
3. Based on our findings and experiments, we have proposed an ontology for Citations' Context and Reasons – *CCRO* that provides abstract conceptualization required to organize citations' relations. *CCRO* has been verified, validated, and assessed by using well-defined procedures and tools proposed in the literature for ontology evaluation. The results show that the proposed ontology is concise, complete, and consistent.
4. To instantiate and map of ontology properties on real data, we have developed a *Mapping Graph* among the verbs with predicative complements in the English Language, the verbs extracted from the selected corpus using *NLP* and the *CCRO* properties.
5. For instantiation and mapping of *Citation's Context and Reasons Ontology* - *CCRO* and its properties on real data, two approaches are used. First is the human annotation with help of domain and linguist experts whereas the second approach is the automatic mapping of citation text using *Mapping Graph*. Two types of data sets are employed in both experiments. One is publicly available and the second is our corpus of citation sentences. After finding accuracy (85.4% and 96.6%), both annotated corpora are converted into Linked Open Data using *RDF* and are visualized using *iGraph* in *R*.
6. To demonstrate the working of the proposed ontology, two different applications are employed. There is a large volume of available scientific literature in each domain and authors spend a lot of time to search and filter the required information. The first application uses *CCRO*'s annotation on the

existing corpus by enriching the citation graph with meaningful tags and minimizing the time and effort to filter the required information. The second application provides a mechanism to author a research article without destroying the useful semantics that in return can help automatic reasoners to filter the required information.

7. *CCRO*'s first application uses the guidelines provided by Shum [32] to discover the evolutionary path of scholarly activity and answers intelligent queries to provide a valid semantic model for *CCRO*. Five different semantic queries are formulated, results are tabulated and visualized.
8. Our second application relies on enriching scientific publications with explicit rhetorical and argumentation discourse structures, using Citation's Context and Reasons Ontology - *CCRO* by annotating citation texts within \LaTeX files at the time of authoring a research paper by the author himself. Furthermore, embedding these structures within *RDF* Data Store enables the creation of semantic publications that become a foundation artifact for the Semantic Publishing Ecosystem and linked resources become part of the current Web of Data.

8.2 Research Contributions

1. Found Problems in Existing Semantic Representation for Citations
2. Discovered and Evaluated 150+ Citation Reasons and Clustered them in 8 Classes
3. Developed, Evaluated and Validated *CCRO*
4. Generated Mapping Graph for Automatic Instantiation of *CCRO* Properties
5. Manual and Automatic Mapping of *CCRO* Properties on Two Corpora
6. Converted Both Annotated Corpora to Linked Open Data using *RDF*
7. Developed *CCRO* Application for Semantic Querying

8. Developed CCRO L^AT_EX Package for Semantic Authoring and Publishing
9. Enabled Semantically Enriched Authoring in LaTeX
10. Established Citation Knowledge Graph Ecosystem

8.3 Future Work

For the future work, here is the list of some but not limited to, related tasks that we are looking forward to:

1. In this research, automatic mapping of *CCRO* ontology has been performed on two different corpora of citation texts, and the accuracy of our approach is calculated. However, there is a need to test the hypothesis on large scale data. In the future, we will be working on a crawler algorithm that can extract citation texts from various research papers available of different digital libraries and create a semantic citation graph using *CCRO*. Semantic queries on this large scale data can shape to identify the evolutionary paths of a research problem or to find the most influential paper within a particular domain of research.
2. Once automatic mapping of *CCRO* on legacy data can be achieved, the next step in our future research is to visualize the semantic citation graph effectively and efficiently. Effective visualization of semantic citation graph can result in reduced time and effort to search and filter required information and to easily navigate among various researches.
3. “Semantic Scholar”¹ is an AI-based tool to find peer-reviewed research from trusted sources. The tool provides various statistics of a query in terms of “Highly Influenced Papers”, “Cite Background”, “Cite Methods” and “Cite Results”. Though Semantic Scholar’s “Highly Influenced Papers” metric is substantially better as compared to raw citation counts, it is simply the sum

¹<https://www.semanticscholar.org/>

of highly influential citations of an author [122]. *CCRO* and its annotation on a large scale data can provide better results for “Highly Influenced Papers”. Therefore, in the future, we will try to collaborate with “Semantic Scholar” with an AI-based solution for the possible inclusion of *CCRO* Citation Reason within their architecture.

4. In the last decade, the need for a machine-actionable representation of scholarly knowledge has emerged. Semantically rich Knowledge Graphs (*SKG*) are using W3C standards such as OWL and RDF to structure and interlink scholarly knowledge. Some examples include Microsoft Academic Knowledge Graph (*MAKG*), Open Research Knowledge Graph, OpenCitations, and OpenAIRE, etc. *CCRO* can play a vital role to conceptualize scholarly knowledge, extraction of entities and concepts, finding semantic connections between entities, and to explore and measure the impact of research.
5. We are also looking forward to developing a scholarly literature publishing framework, where the complete research paper, along-with its citation texts and their reasons, can be stored in the form of *Linked Open Data*. Thus, documenting future research in a more meaningful way.
6. Last but not least, combining the two (2) systems; one that can discover citation reasons from existing scholarly data and the other that can be used by authors to integrate these reasons at the time of authoring a research article, will become *Semantic Enabled Publishing Framework*.

Bibliography

- [1] S. B. Shum, “Evolving the web for scientific knowledge: First step towards an ‘HCI knowledge web’,” *Interfaces, British HCI Group Magazine*, vol. 39, pp. 16–21, 1998.
- [2] S. Teufel, “Argumentative Zoning : Information Extraction from Scientific Text,” *PhD Dissertation, University of Edinburgh*, 1999. [Online]. Available: <https://www.cl.cam.ac.uk/~sht25/thesis/t1.pdf>
- [3] L. M. Baird and C. Oppenheim, “Do citations matter?” *Journal of Information Science*, vol. 20, no. 1, pp. 2–15, 1994.
- [4] H. Small, “Interpreting maps of science using citation context sentiments: A preliminary investigation,” *Scientometrics*, vol. 87, no. 2, pp. 373–388, 2011.
- [5] E. Garfield, “Citation Analysis as a Tool in Journal Evaluation: Journals can be ranked by frequency and impact of citations for science policy studies,” *Science*, vol. 178, no. 4060, pp. 471–479, 1972.
- [6] B. Cronin, “The Need for a Theory of Citing,” *Journal of Documentation*, vol. 37, no. 1, pp. 16–24, 1981.
- [7] L. Leydesdorff, “Theories of citation?” *Scientometrics*, vol. 43, no. 1, pp. 5–25, 1998.
- [8] B. Yu, *Automated Citation Sentiment Analysis: What Can We Learn from Biomedical Researchers*. American Society for Information Science, 2013.
- [9] P. L. Phugnar, “A citation analysis of doctoral dissertation in library and information science accepted by the universities in Western India,”

- PhD Dissertation, Tilak Maharashtra Vidyapeeth, Pune, 2012.* [Online]. Available: <http://hdl.handle.net/10603/18612>
- [10] V. Wilson, “Research Methods: Bibliometrics,” *Evidence Based Library and Information Practice*, vol. 7, no. 3, pp. 121–123, 2012.
- [11] Y. Ding, G. Zhang, T. Chambers, M. Song, X. Wang, and C. Zhai, “Content-based citation analysis: The next generation of citation analysis,” *Journal of the Association for Information Science and Technology*, vol. 65, no. 9, pp. 1820–1833, 2014.
- [12] I. G. Councill, C. L. Giles, and M.-y. Kan, “ParsCit: An open-source CRF Reference String Parsing Package,” in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*. European Language Resources Association (ELRA), 2008, pp. 661–667.
- [13] J. Markoff, G. Shapiro, and S. R. Weitrnan, “Toward the Integration of Content Analysis and General Methodology,” *Sociological Methodology*, vol. 5, pp. 1–58, 1975.
- [14] G. N. Gilbert and S. Woolgar, “The quantitative study of science: An examination of the literature,” *Social Studies of Science*, vol. 4, no. 3, pp. 279–294, 1974.
- [15] J. Xu, Y. Zhang, Y. Wu, J. Wang, X. Dong, and H. Xu, “Citation Sentiment Analysis in Clinical Trial Papers,” in *Annual Symposium Proceedings*. AMIA, 2015, pp. 1334–1341.
- [16] B. H. Butt, M. Rafi, A. Jamal, R. S. Ur Rehman, S. M. Z. Alam, and M. B. Alam, “Classification of research citations (CRC),” in *CEUR Workshop Proceedings*. CEUR, 2015, pp. 18–27.
- [17] I. C. Kim and G. R. Thoma, “Automated classification of author’s sentiments in citation using machine learning techniques: A preliminary study,” in *Proceedings of the IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB*. IEEE, 2015, pp. 1–7.

- [18] P. A. H. Kazi and M. S. Patwardhan, "Context based citation summary of research articles: A step towards qualitative citation index," in *Proceedings of the IEEE International Conference on Computer Communication and Control, IC4*. IEEE, 2016, pp. 1–6.
- [19] M. H. Alvarez, J. M. Soriano, and P. Martínez-Barco, "Citation function, polarity and influence classification," *Natural Language Engineering*, vol. 23, no. 4, pp. 561–588, 2017.
- [20] Z. Taskin and U. Al, "A content-based citation analysis study based on text categorization," *Scientometrics*, vol. 114, pp. 335–357, 2017.
- [21] K. Halil, P. Zeshan, T. Shabnam, T. Tung, R. Graciela, and S. Jodi, "Confirm or refute?: A comparative study on citation sentiment classification in clinical research publications," *Journal of Biomedical Informatics*, vol. 91, pp. 103–123, 2019.
- [22] E. Orduña-Malea, J. M. Ayllón, A. Martín-Martín, and E. Delgado López-Cózar, "About the size of Google Scholar: playing the numbers," *arXiv preprint arXiv:1407.6239*, 2014.
- [23] B. Motik, B. Parsia, and P. F. Patel-Schneider, *OWL 2 Web Ontology Language XML Serialization (Second Edition)*. W3C Recommendation, 2012.
- [24] S. Peroni and D. Shotton, "The SPAR Ontologies," in *Proceedings of The Semantic Web – ISWC 2018*. Springer International Publishing, 2018, pp. 119–136.
- [25] —, "FaBiO and CiTO: Ontologies for describing bibliographic resources and citations," *Journal of Web Semantics*, vol. 17, pp. 33–43, 2012.
- [26] P. Ciancarini, A. Di Iorio, A. G. Nuzzolese, S. Peroni, and F. Vitali, "Evaluating citation functions in CiTO: Cognitive issues," *Lecture Notes in Computer Science*, vol. 8465, no. LNCS, pp. 580–594, 2014.

- [27] A.-M. Boanerges, H. Farshad, A. I. Budak, and S. Amit P., “SwetoDblp ontology of Computer Science publications,” *Journal of Web Semantics*, vol. 5, no. 3, pp. 151–155, 2007.
- [28] P. Silvio, D. Alexander, G. Tanya, and S. David, “Setting our bibliographic references free: towards open citation data,” *Journal of Documentation*, vol. 71, no. 2, pp. 253–277, 2015.
- [29] W. Ruijie, Y. Yuchen, W. Jialu, J. Yuting, Z. Ye, Z. Weinan, and W. Xinbing, “AceKG: A Large-scale Knowledge Graph for Academic Data Mining,” in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management. CIKM’18*, 2018, pp. 1487–1490.
- [30] F. Michael, “The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data,” *The Semantic Web - ISWC*, vol. 11779, pp. 113–129, 2019.
- [31] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, and A. Kanakia, “Microsoft Academic Graph: When experts are not enough,” *Quantitative Science Studies*, vol. 1, no. 1, pp. 396–413, 2020.
- [32] S. B. Shum, T. Clark, A. de Waard, T. Groza, S. Handschuh, and A. Sandor, “Scientific Discourse on the Semantic Web: A Survey of Models and Enabling Technologies,” *Semantic Web Journal: Interoperability, Usability, Applicability*, vol. Special Issue on Survey Articles, 2010.
- [33] T. Groza, S. Handschuh, K. Möller, and S. Decker, “SALT - Semantically Annotated LaTeX for Scientific Publications,” *The Semantic Web: Research and Applications. Lecture Notes in Computer Science*, vol. 4519, pp. 518–532, 2007.
- [34] A. Di Iorio, A. G. Nuzzolese, S. Peroni, D. Shotton, and F. Vitali, “Describing bibliographic references in RDF,” in *Proceedings of 4th Workshop on Semantic Publishing - SePublica*, vol. 1155. CEUR, 2014, pp. 41–56.

- [35] B. Tillett, "What is FRBR? A conceptual model for the bibliographic universe," *Australian Library Journal*, vol. 54, no. 1, pp. 24–30, 2005.
- [36] A. Constantin, S. Peroni, S. Pettifer, D. Shotton, and F. Vitali, "The Document Components Ontology (DoCO)," *Semantic Web*, vol. 7, no. 2, pp. 167–181, 2016.
- [37] P. Ciccarese, E. Wu, G. Wong, M. Ocana, J. Kinoshita, A. Ruttenberg, and T. Clark, "The SWAN biomedical discourse ontology," *Journal of Biomedical Informatics*, vol. 41, no. 5, pp. 739–751, 2008.
- [38] C. Biemann, "Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems," in *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*. Association for Computational Linguistics, 2006, pp. 73–80.
- [39] S. Teufel, a. Siddharthan, and D. Tidhar, "An annotation scheme for citation function," in *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*. ACM, 2006, pp. 80–87.
- [40] E. Garfield, "Citation frequency as a measure of research activity and performance," *Essay of an Information Scientist*, vol. 1, pp. 406–408, 1973.
- [41] M. J. Moravcsik and P. Murugesan, "Some Results on the Function and Quality of Citations," *Social Studies of Science*, vol. 5, no. 1, pp. 86–92, 1975.
- [42] E. Garfield, "When to Cite," *The Library Quarterly: Information, Community, Policy*, vol. 66, no. 4, pp. 449–458, 1996.
- [43] S. Teufel, a. Siddharthan, and D. Tidhar, "Automatic classification of citation function," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. ACM, 2006, pp. 103–110.
- [44] A. Athar and S. Teufel, "Context-Enhanced Citation Sentiment Detection," in *Proceedings of the 2012 Conference of the North American Chapter of the*

- Association for Computational Linguistics: Human Language Technologies*.
ACM, 2012, pp. 597–601.
- [45] A. Athar, “Sentiment analysis of citations using sentence structure-based features,” in *Proceedings of the ACL 2011 Student Session*. ACL, 2011, pp. 81–87.
- [46] C. Dong and U. Schäfer, “Ensemble-style Self-training on Citation Classification,” in *Proceedings of 5th International Joint Conference on Natural Language Processing*. ACL, 2011, pp. 623–631.
- [47] X. Li, Y. He, A. Meyers, and R. Grishman, “Towards Fine-grained Citation Function Classification,” *Ranlp*, pp. 402–407, 2013.
- [48] M. Valenzuela, V. Ha, and O. Etzioni, “Identifying Meaningful Citations,” in *Proceedings of the AAAI Workshop on Scholarly Big Data*. AAAI, 2015, p. 6.
- [49] M. a. Angrosh, S. Cranefield, and N. Stanger, “Context identification of sentences in related work sections using a conditional random field: towards intelligent digital libraries,” in *Proceedings of the 10th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM/IEEE, 2010, pp. 293–302.
- [50] N. Tandon and A. Jain, “Citation Context Sentiment Analysis for Structured Summarization of Research Papers,” in *Proceedings of the 35th German Conference on Artificial Intelligence (KI-2012)*. Springer, 2012, pp. 98–102.
- [51] C. Jochim and H. Schütze, “Towards a Generic and Flexible Citation Classifier Based on a Faceted Classification Scheme,” in *Proceedings of COLING’12*. ACL, 2012, pp. 1343–1358.
- [52] E. M. Han Xu and A. Mahidadia, “Using Heterogeneous Features for Scientific Citation Classification,” in *Proceedings of the 13th Conference of the Pacific Association for Computational Linguistics*. Sig Media, 2013, pp. 1–8.

- [53] X. Wan and F. Liu, “Are all literature citations equally important? Automatic citation strength estimation and its applications,” *Journal of the Association for Information Science and Technology*, vol. 65, no. 9, pp. 1929–1938, 2014.
- [54] M. Hernández A. and J. M. Gómez, “Survey in sentiment, polarity and function analysis of citation,” in *Proceedings of the First Workshop on Argumentation Mining*. ACL, 2014, pp. 102–103.
- [55] D. Jurgens, S. Kumar, R. Hoover, D. McFarland, and D. Jurafsky, “Measuring the Evolution of a Scientific Field through Citation Frames,” *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 391–406, 2018.
- [56] B. Khadidja, N. Zhendong, and N. Ally S., “A Preliminary Study to Compare Deep Learning with Rule-based Approaches for Citation Classification,” in *Proceedings of International Conference on Control, Artificial Intelligence, Robotics Optimization (ICCAIRO)*, 2018, pp. 43–47.
- [57] J. Meng, “Citation Function and Polarity Classification in Biomedical Papers,” *MS Thesis, The University of Western Ontario*, 2018. [Online]. Available: <https://ir.lib.uwo.ca/etd/5367/>
- [58] A. Cohan, W. Ammar, M. van Zuylen, and F. Cady, “Structural Scaffolds for Citation Intent Classification in Scientific Publications,” in *NAACL-HLT*, 2019, p. 11.
- [59] F. Qayyum and M. Afzal, “Identification of important citations by exploiting research articles’ metadata and cue-terms from content,” *Scientometrics*, vol. 118, no. 1, pp. 21–43, 2019.
- [60] J. Perier-Camby, M. Bertin, I. Atanassova, and F. Armetta, “A Preliminary Study to Compare Deep Learning with Rule-based Approaches for Citation Classification,” in *Proceedings of Workshop on Bibliometric-enhanced Information Retrieval - BIR*, 2019, pp. 125–131.

- [61] H. Zhao, L. Zhunchen, C. Feng, and Y. Yuming, “A Context-based Framework for Resource Citation Classification in Scientific Literatures,” in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 1041–1044.
- [62] T. Suppawong, K. Sung Woo, W. Poom, P. Chanatip, S. Tanakitti, S.-U. Hassan, and H. Peter, “Automatic Classification of Algorithm Citation Functions in Scientific Literature,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 10, pp. 1881–1896, 2020.
- [63] W. Mingyang, Z. Jiaqi, J. Shijia, Z. Xiangrong, Z. Na, and G. Chen, “Important citation identification by exploiting the syntactic and contextual information of citations,” *Scientometrics*, vol. 125, p. 21, 2020.
- [64] R. M. Shiffrin and K. Börner, “Mapping Knowledge Domains,” in *Proceedings of the National Academy of Sciences*. PNAS, 2004, pp. 5183–5185.
- [65] C. Mancini and S. J. Buckingham Shum, “Modelling discourse in contested domains: A semiotic and cognitive framework,” *International Journal of Human Computer Studies*, vol. 64, no. 11, pp. 1154–1171, 2006.
- [66] S. B. Shum, E. Motta, and J. Domingue, “ScholOnto: An ontology-based digital library server for research documents and discourse,” *International Journal on Digital Libraries*, vol. 3, no. 3, pp. 237–248, 2000.
- [67] M. Clara and S. B. Shum, “Modelling discourse in contested domains: A semiotic and cognitive framework,” *International Journal of Human Computer Studies*, vol. 64, no. 11, pp. 1154–1171, 2006.
- [68] A. Di Iorio, A. G. Nuzzolese, F. Osborne, S. Peroni, F. Poggi, M. Smith, F. Vitali, and J. Zhao, “The RASH Framework: enabling HTML+RDF submissions in scholarly venues,” in *Proceedings of the 14th International Semantic Web Conferences (ISWC)*, 2015, p. 4.
- [69] S. Peroni, F. Osborne, A. Di Iorio, A. G. Nuzzolese, F. Poggi, F. Vitali, and E. Motta, “Research Articles in Simplified HTML: a Web-first format for

- HTML-based scholarly articles,” *PeerJ Computer Science*, vol. 3, p. e132, 2017.
- [70] S. Ait-Mokhtar, J. P. Chanod, and C. Roux, “Robustness beyond shallowness: Incremental deep parsing,” *Natural Language Engineering*, vol. 8, no. 3, pp. 121–144, 2002.
- [71] P. Ciccarese, M. Ocana, L. J. Garcia Castro, S. Das, and T. Clark, “An open annotation ontology for science on web 3.0,” *Journal of Biomedical Semantics*, vol. 2, no. Suppl 2, 2011.
- [72] D. Arseneau, “The cite package: well formed numeric citations,” *TeXdoc Online*. [Online]. Available: <http://texdoc.net/texmf-dist/doc/latex/cite/cite.pdf>
- [73] F. Garcia, “LATEX and the different bibliography styles,” *The PracTEX Journal*, vol. 28, no. 2, 2007.
- [74] P. Williams and T. Schnier, “The Harvard Family of Bibliography Styles,” 1994.
- [75] M. Swift, “The achicago LaTeX package - Chicago Manual author-date citations,” 2001.
- [76] P. W. Daly, “Natural Sciences Citations and References (Author - Year and Numerical Schemes),” *Sciences New York*, 2009.
- [77] P. Daly, “Natural Sciences Citations and References,” *Texdoc.Net*, 2010.
- [78] E. Meijer, “The apacite package,” *English*, 2009.
- [79] A. Gómez-Pérez, M. Fernández-López, and O. Chorco, *Ontological Engineering: With Examples From the Areas of Knowledge Management, E-Commerce and the Semantic Web*. Springer, 2006.
- [80] M. Fernández-López, A. Gómez-Pérez, and N. Juristo, “METHONTOLOGY: From Ontological Art Towards Ontological Engineering,” in *AAAI Symposium on Ontological Engineering*. Stanford, 1997, pp. 33–40.

- [81] B. Levin, *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, 1993.
- [82] N. S. Young, J. P. Ioannidis, and O. Al-Ubaydli, “Why current publication practices may distort science,” *PLoS Medicine*, vol. 5, no. 10, pp. 1418–1422, 2008.
- [83] G. Zhang, Y. Ding, and S. Milojević, “Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content,” *Journal of the American Society for Information Science and Technology*, vol. 64, no. 7, pp. 1490–1503, 2013.
- [84] D. R. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara, “The ACL anthology network corpus,” *Language Resources and Evaluation*, vol. 47, no. 4, pp. 919–944, 2013.
- [85] A. Athar, “Sentiment analysis of scientific citation,” *University of Cambridge, Computer Laboratory*, vol. UCAM-CL-TR, no. 856, p. 114, 2014.
- [86] K. Hyland, *Hedging in Scientific Research Articles*. John Benjamins Publishing Company, 1998.
- [87] G. Thomson and Y. Yiyun, “Evaluation in the Reporting Verbs Used in Academic Papers,” *Applied Linguistics*, vol. 12, no. 4, pp. 365–382, 1991.
- [88] —, “Emergent Grammar,” *Berkeley Linguistics Society*, vol. 13, pp. 139–157, 1987.
- [89] B. Yu, “Automated citation sentiment analysis: What can we learn from biomedical researchers,” *Proceedings of the Association for Information Science and Technology*, vol. 50, no. 1, pp. 1–9, 2013.
- [90] J. R. Finkel, T. Grenager, and C. Manning, “Incorporating non-local information into information extraction systems by Gibbs sampling,” in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. ACL, 2005, pp. 363–370.

- [91] P. Slav, “Announcing SyntaxNet: The World’s Most Accurate Parser Goes Open Source,” *Google AI Blog*, 2016. [Online]. Available: <https://ai.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html>
- [92] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, “The Stanford CoreNLP Natural Language Processing Toolkit,” in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. ACL, 2014, pp. 55–60.
- [93] N. A. Manan and N. M. Noor, “Analysis of Reporting Verbs in Master’s Theses,” *Procedia - Social and Behavioral Sciences*, vol. 134, no. 15, pp. 140–145, 2014.
- [94] P. Alberto, “How many scholarly articles are written in LaTeX?” *Authorea Inc.*, 2016. [Online]. Available: <https://doi.org/10.22541/2Fau.148771883.35456290>
- [95] S. G. Kolte and S. G. Bhirud, “WordNet: A Knowledge Source for Word Sense Disambiguation,” *International Journal of Recent Trends in Engineering*, vol. 2, no. 4, pp. 213–217, 2009.
- [96] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, *The Description Logic Handbook: Theory, Implementation and Applications - 2nd Edition*. Cambridge University Press, 2010.
- [97] B.-A. Lipetz, “Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators,” *American Documentation*, vol. 16, no. 2, pp. 81–90, 1965.
- [98] D. E. Chubin and S. D. Moitra, “Content Analysis of References: Adjunct or Alternative to Citation Counting?” *Social Studies of Science*, vol. 5, no. 4, pp. 423–441, 1975.
- [99] I. Spiegel-Rosing, “Science Studies: Bibliometric and Content Analysis,” *Social Studies of Science*, vol. 7, no. 1, pp. 97–113, 1977.

- [100] C. O. Frost, "Use of citations in literary research - preliminary classification of citation functions," *The Library Quarterly: Information, Community, Policy*, vol. 49, no. 4, pp. 339–414, 1979.
- [101] C. Oppenheim and E. Garfield, "Highly cited old papers," *Journal of the American Society for Information Science*, vol. 31, no. 3, pp. 219–219, 1980.
- [102] B. C. Peritz, "A classification of citation roles for the social sciences and related fields," *Scientometrics*, vol. 5, no. 5, pp. 303–312, 1983.
- [103] K. W. McCain and K. Turner, "Citation context analysis and aging patterns of journal articles in molecular genetics," *Scientometrics*, vol. 17, no. 1-2, pp. 127–163, 1989.
- [104] B. Jörg, "Towards the Nature of Citations," in *Poster Proceedings of the 5th International Conference on Formal Ontology in Information Systems*. IOS, 2008, pp. 31–36.
- [105] R. Jha, A. A. Jbara, V. Qazvinian, and D. R. Radev, "NLP-driven citation analysis for scientometrics," *Natural Language Engineering*, vol. 23, no. 1, pp. 93–130, 2017.
- [106] B. Glimm, I. Horrocks, B. Motik, G. Stoilos, and Z. Wang, "HermiT: An OWL 2 Reasoner," *Journal of Automated Reasoning*, vol. 53, no. 3, pp. 245–269, 2014.
- [107] M. Fahad, M. A. Qadir, and M. W. Noshairwan, "Ontological errors Inconsistency, Incompleteness and Redundancy," in *Proceedings of the 10th International Conference on Enterprise Information Systems*. ISAS, 2008, pp. 253–258.
- [108] I. Ihsan, S. Imran, and O. Ahmed, "Sentiment Based Study of Citations' Reporting Verb Corpus Using Natural Language Processing," *Journal of Corpus Linguistics*, vol. 1, no. 2, p. In Press, 2018.

- [109] M. Charles, “Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines,” *English for Specific Purposes*, vol. 25, no. 3, pp. 310–331, 2006.
- [110] S. Shekarpour, V. Shalin, K. Thirunarayan, and A. P. Sheth, “CEVO: Comprehensive Event Ontology Enhancing Cognitive Annotation,” *CoRR*, vol. abs/1701.0, 2017. [Online]. Available: <http://arxiv.org/abs/1701.05625>
- [111] R. Artstein and M. Poesio, “Inter-Coder Agreement for Computational Linguistics,” *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, 2008.
- [112] S. Babar and P. D. Patil, “Improving Performance of Text Summarization,” *Procedia Computer Science*, vol. 46, pp. 354–363, 2015.
- [113] J. Brownlee, “Tour of Evaluation Metrics for Imbalanced Classification,” *Machine Learning Mastery*, 2020. [Online]. Available: <https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/>
- [114] M. Bertin, I. Atanassova, C. R. Sugimoto, and V. Lariviere, “The linguistic patterns and rhetorical structure of citation context: an approach using n-grams,” *Scientometrics*, vol. 109, no. 3, pp. 1417–1434, 2016.
- [115] R. E. Mercer and C. Di Marco, “The Importance of Fine-Grained Cue Phrases in Scientific Citations,” *Advances in Artificial Intelligence*, vol. 2671, pp. 550–556, 2003.
- [116] L. Galke, T. Melnychuk, E. Seidlmayer, S. Trog, K. U. Förstner, C. Schultz, and K. Tochtermann, “Inductive Learning of Concept Representations from Library-Scale Corpora with Graph Convolution,” *Lecture Notes in Informatics (LNI) - Proceedings*, vol. P-294, pp. 219–232, 2019.
- [117] J. E. Hirsch, “An index to quantify an individual’s scientific research output,” *Proceedings of the National Academy of Sciences, U.S.A.*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

-
- [118] K. R. Dienes, “Completing h,” *Journal of Informetrics*, vol. 9, no. 2, pp. 385–397, 2015.
- [119] F. Brischoux and P. Legagneux, “Don’t Format Manuscripts,” *Scientist (Philadelphia, Pa.)*, vol. 23, pp. 24–24, 2009.
- [120] S. Pakin, “How to Package Your LATEX Package,” *Texdoc.Net*, 2015. [Online]. Available: <http://texdoc.net/texmf-dist/doc/latex/dtxtut/dtxtut.pdf>
- [121] T. Berners-Lee, J. Hendler, and O. Lassila, “The semantic web,” *Scientific American*, vol. 284, no. 5, pp. 34–43, 2001.
- [122] O. Etzioni, “AI zooms in on highly influential citations,” *Nature*, vol. 547, pp. 32–32, 2009.

Appendix A:

Citation Reasons

TABLE 1: Identified Citation Reason

Sr.	Source	Reasons	Year	Citation Reasons
1.	Liptez [97]	4	1965	Scientific Contribution, Continuity Relationship, Disposition of Contribution, Non-Scientific Contribution
2.	Chubin & Moitra [98]	6	1975	Affirmative Perfunctory, Affirmative Subsidiary, Negative Partial, Negative Total, Affirmative Basic, Affirmative Additional
3.	Moravesik & Murugesan [41]	4	1975	Conceptual, Evolutionary, Perfunctory, Confirmative
4.	Spiegel-Rosing [99]	7	1977	Concept, Point of Departure, Comparative, History, Data in Text, Data in Tables, Interpretation
5.	Frost [100]	5	1979	Factual Evidence, Primary Text, View of Other Scholars, Further Reading, Previous Scholarship
6.	Oppenheim & Garfield [101]	7	1980	Theoretical Equation, Theory Not Applicable, Historical Background, Data (Comparative), Data (Not Comparative), Relevant Work, Methodology
7.	Pertiz [102]	6	1983	Argumentative, Setting Stage, Background, Comparative, Documentary, Methodology

TABLE 1: Identified Citation Reason (Continued)

Sr.	Source	Reasons	Year	Citation Reasons
8.	McCain & Turner [103]	5	1989	Results Central, Results Peripheral, Introduction-Central, Introduction-Peripheral, Methods-Central
9.	Eugene Garfield [42]	15	1996	Giving Homage, Giving Credit, Sustaining Claims, Correct Own Work, Correct Others Work, Alert Forthcoming Work, Authenticate Data, Identify Idea, Identify Concept, Criticize, Disclaim Work, Dispute Priority, Identify Methodology, Background, Lead to Uncited Work
10.	Teufel, Siddharthan, & Tidhar [39]	7	2006	Alternate Approach, Research Gap, Current Work, Background, Introduction, Citation Sentences, Descriptive Sentences
11.	Teufel, Siddharthan, & Tidhar [43]	4	2006	Positive, Contrast, Weakness, Neutral
12.	Jörg [104]	6	2008	Inspired, Extended, Based, Proposed, Compare, Described
13.	Angrosh, Cranefield, & Stanger [49]	4	2010	Citation Positive Sentiment, Contrast, Weakness, Neutral
14.	Athar [45]	3	2011	Positive, Negative, Neutral
15.	Dong and Schäfer [46]	3	2011	Fundamental Idea, Technical Basis, Background
16.	Athar and Teufel [44]	3	2012	Positive, Negative, Neutral
17.	Tandon and Jain [50]	5	2012	Strength, Application, Limitations, Summary, Related Work
18.	Jochim & Schütze [51]	3	2012	Positive, Negative, Neutral
19.	Yu [89]	5	2013	Neutral (Implicitly Positive), Positive, Negative, Mitigated, Neutral
20.	Li, He, Meyers, & Grishman [47]	12	2013	Based on, Corroboration, Discover, Positive, Practical, Significant, Standard, Supply, Negative, Contrast, Cocitation, Neutral

TABLE 1: Identified Citation Reason (Continued)

Sr.	Source	Reasons	Year	Citation Reasons
21.	Han Xu, Eric Martin [52]	3	2013	Functional, Perfunctory, Hard to Tell
22.	Hernández A. & Gómez [54]	7	2014	Based On, Useful, Correct, Debate, Contrast, Weakness, Acknowledge
23.	Wan and Liu [53]	2	2014	Strength, Importance
24.	Valenzuela, Ha, & Etzioni [48]	4	2015	Using the Work, Extending the Work, Comparison, Related Work
25.	Butt [16]	2	2015	Positive, Negative
26.	Kim and Thoma [17]	3	2015	Positive, Negative, Neutral
27.	Xu [15]	3	2015	Positive, Negative, Neutral
28.	Kazi and Patwardan [18]	3	2016	Positive, Negative, Self-Citation
29.	Jha, Jbara, Qazvinian, and Radev [105]	6	2017	Use, Sub-stain, Basis, Criticize, Comparison, Neutral
30.	Taşkın and Al [20]	5	2017	Method, Data Validation, Literature Review, Definition, Data
31.	Alvarez et al. [19]	8	2017	Acknowledgment, Corroboration, Based On, Supply, Useful, Weakness, Hedges, Comparison (Contrast)
32.	Jurgens et al. [55]	7	2018	Background, Motivation, Uses, Extension, Continuation, Comparison/Contrast, Future
33.	Khadidja et al. [56]	6	2018	Based On, Useful, Acknowledge, Contrast, Weakness, Hedges
34.	Jia Meng [57]	8	2018	Background, Statement, Comparison, Multi-Comparison, Confirmation, Being Confirmed, Contrast/Conflict, Unsolved
35.	Cohan et al. [58]	3	2019	Background, Method, Result Comparison

TABLE 1: Identified Citation Reason (Continued)

Sr.	Source	Reasons	Year	Citation Reasons
36.	Qayyum & Afzal [59]	2	2019	Important, Not Important
37.	Perier-Camby et al. [60]	7	2019	Background, Motivation, Uses, Extension, Continuation, Comparison/Contrast, Future
38.	Halil et al. [21]	3	2019	Positive, Negative, Neutral
39.	Zhao et al. [61]	6	2019	Produce, Use, Introduce, Extent, Compare, Other
40.	Suppawong et al. [62]	4	2020	Use, Extend, Mention, Notalgo
41.	Mingyang et al. [63]	3	2020	Comparison, Utilize the Work, Extend the Work

Appendix B:

Reporting Verbs

Each citation text of *AAN* dataset was read by a POS Tagger and the verbs were tagged in 6 different verb phrases. A complete set of 8700 citation texts with tagged verb phrases is pre-processed using stemming and lemmatization alternatively. Once the tagging is complete, a simple algorithm finds the unique verbs and calculates their frequency (No of times a unique verb phrase used in the complete dataset). If a verb phrase has high frequency, it means that it is commonly used verb while citing, thus carries more weight in the corpus. Frequency of a verb can play a vital role in identifying a reporting verb. Following is a list of high occurring verbs termed as *Reporting Verbs*.

TABLE 2: Reporting Verbs

Sr.	Verb	Sr	Verb	Sr	Verb	Sr.	Verb
1.	use	21.	introduce	41.	derive	61.	discuss
2.	propose	22.	estimate	42.	allow	62.	reduce
3.	describe	23.	take	43.	maximize	63.	explore
4.	include	24.	produce	44.	select	64.	modify
5.	show	25.	define	45.	assign	65.	model
6.	train	26.	develop	46.	label	66.	exploit
7.	give	27.	consider	47.	classify	67.	depend
8.	follow	28.	make	48.	implement	68.	rely
9.	see	29.	measure	49.	require	69.	cluster
10.	provide	30.	achieve	50.	refer	70.	indicate
11.	compare	31.	employ	51.	focus	71.	involve
12.	compute	32.	learn	52.	annotate	72.	demonstrate
13.	perform	33.	determine	53.	relate	73.	examine
14.	obtain	34.	find	54.	suggest	74.	prove
15.	improve	35.	parse	55.	test	75.	correlate
16.	evaluate	36.	build	56.	calculate	76.	argue
17.	report	37.	apply	57.	extend	77.	agree
18.	present	38.	combine	58.	choose	78.	outperform
19.	generate	39.	represent	59.	work	79.	fail
20.	tag	40.	identify	60.	incorporate	80.	omit

Appendix C:

Citations' Context and Reasons

Ontology

IRI:	http://ccropus/ccro
Version IRI:	http://ccropus/ccro/2019-10-20
Current Version:	1.4.3
Authors:	Imran Ihsan, M. Abdul Qadir
Imported Ontologies:	OLiA ² - Ontologies of Linguistic Annotation

Abstract

The Citations' Context and Reasons Ontology (CCRO) is an ontology that provides abstract conceptualization required to organize citations' relations. Each reason in the ontology defines a unique citation link between research papers in a citation graph.

CCRO defines a taxonomical hierarchy of eight object properties distributed among three main sentiment-based reasons. Using the ontology concept, a citation can

²<http://nachhalt.sfb632.uni-potsdam.de/owl/olia.owl>

have ‘Positive’, ‘Negative’ or ‘Neutral’ context or sentiment with a possibility of being a part of one of its constituent properties.

Classes

Paper, CitingPaper, CitedPaper, Citation, MainVerb
--

Paper

IRI: <http://ccropus/ccro/paper>

Any Research Paper.

has sub-class: CitingPaper, CitedPaper

CitingPaper

IRI: <http://ccropus/ccro/citingpaper>

A Research Paper that cites other research papers.

has super-class: Paper

CitedPaper

IRI: <http://ccropus/ccro/citedpaper>

A Research Paper that is cited by others.

has super-class: Paper

Citation

IRI: <http://ccropus/ccro/citation>

Citation Sentence in a Citing Paper.

MainVerb

IRI: <http://ccropus/ccro/citation>

Citation Sentence in a Citing Paper.

is equivalent-to: OLiA:MainVerb
--

Object Properties

consistsOf, isStructuredAround, Reason, citesWithPositiveReason, citesWithNegativeReason, citesWithNeutralReason, Incorporate, Extend, BasedOn, Negate, Criticize, Contrast, Compare, Discuss

ccro:consistsOf

IRI: <http://ccropus/ccro/consistsOf>

The property specifies a citation texts within a citing paper.

has domain: CitingPaper

has range: Citation

ccro:isStructuredAround

IRI: <http://ccropus/ccro/isStructuredAround>

The property defines the main or reporting verb within a citation text.

has domain: Citation

has range: MainVerb

ccro:Reason

IRI: <http://ccropus/ccro/Reason>

The property defines any arbitrary reason for citation between the citing and cited paper.

has domain: CitingPaper

has range: CitedPaper

has sub-properties: citesWithPositiveReason, citesWithNegativeReason, citesWithNeutralReason

ccro:citesWithPositiveReason

IRI: <http://ccropus/ccro/citesWithPositiveReason>

The property defines reason for citation with positive sentiment between the citing and cited paper.

has domain: CitingPaper

has range: CitedPaper

has super-property: Reason

has sub-properties: Incorporate, Extend, BasedOn

disjoint-with: citesWithNegativeReason

ccro:citesWithNegativeReason

IRI: <http://ccropus/ccro/citesWithNegativeReason>

The property defines reason for citation with negative sentiment between the citing and cited paper.

<p>has domain: CitingPaper</p> <p>has range: CitedPaper</p> <p>has super-property: Reason</p> <p>has sub-properties: Negate, Criticize, Contrast</p> <p>disjoint-with: citesWithPositiveReason</p>

ccro:citesWithNeutralReason

IRI: <http://ccropus/ccro/citesWithNeutralReason>

The property defines reason for citation with neutral sentiment between the citing and cited paper.

<p>has domain: CitingPaper</p> <p>has range: CitedPaper</p> <p>has super-property: Reason</p> <p>has sub-properties: Compare, Discuss</p>

ccro:Incorporate

IRI: <http://ccropus/ccro/Incorporate>

To cite a research as part of a whole.

<p>has domain: CitingPaper</p> <p>has range: CitedPaper</p> <p>has super-property: citesWithPositiveReason</p>

ccro:Extend

IRI: <http://ccropus/ccro/Extend>

To spread from a central research to a wider solution.

<p>has domain: CitingPaper</p> <p>has range: CitedPaper</p> <p>has super-property: citesWithPositiveReason</p>

ccro:BasedOn

IRI: <http://ccropus/ccro/BasedOn>

To use a research as foundation or starting point.

has domain: CitingPaper
has range: CitedPaper
has super-property: citesWithPositiveReason

ccro:Negate

IRI: <http://ccropus/ccro/Negate>

To cause to be ineffective or invalid.

has domain: CitingPaper
has range: CitedPaper
has super-property: citesWithNegativeReason

ccro:Criticize

IRI: <http://ccropus/ccro/Criticize>

To find fault in a research with: points out the faults of.

has domain: CitingPaper
has range: CitedPaper
has super-property: citesWithNegativeReason

ccro:Contrast

IRI: <http://ccropus/ccro/Contrast>

To show differences with opposite nature.

<p>has domain: CitingPaper</p> <p>has range: CitedPaper</p> <p>has super-property: citesWithNegativeReason</p>

ccro:Compare

IRI: <http://ccropus/ccro/Compare>

To examine in order to show similarities.

<p>has domain: CitingPaper</p> <p>has range: CitedPaper</p> <p>has super-property: citesWithNeutralReason</p>
--

ccro:Discuss

IRI: <http://ccropus/ccro/Discuss>

To consider or examine by argument.

<p>has domain: CitingPaper</p> <p>has range: CitedPaper</p> <p>has super-property: citesWithNeutralReason</p>
--

Appendix D:

Questionnaire for User-based Evaluation of CCRO

For a user-based evaluation of *CCRO*, a questionnaire has been designed using the guidelines provided by Gómez et al. [79] and Fahad et al. [107]. In general questionnaire contains five (05) questions for “Incomplete Concept Classification”, “Disjoint Knowledge Omission”, “Exhaustive Knowledge Omission”, “Scientific Knowledge Omission” and “Redundancy of Disjoint Relations”. Each questions contains, an initial binary “Yes or No” answer followed by a descriptive section to explain the choice made. Domain experts with vast experience in Digital Library, NLP and Computational Linguistics have evaluated the ontology. Table 4.5 outlines the questionnaire used to evaluate *CCRO*.

TABLE 3: Questionnaire for User-based Evaluation of CCRO

Question	Binary
1. Incomplete Concept Classification	
Do you think that any citation reason or context is missing from the ontology?	Yes, No
If Yes , list missing citation reason.	
<i>Hint: Check for the errors if CCRO has overlooked any citation reason that is present in the related domain while classifying reasons.</i>	
2. Disjoint Knowledge Omission	
Is there any disjoint relationship missing from the proposed ontology?	Yes, No
If Yes , describe pair that exhibits disjoint relationship.	
<i>Hint: Check for the partitions or sub-classes where CCRO has omitted a disjoint knowledge axiom between citation reason's classes or properties.</i>	
3. Exhaustive Knowledge Omission	
Is the ontology describes exhaustive sub-classes for its constituent concepts?	Yes, No
If Yes , describe concepts and their sub-classes that are missing.	
<i>Hint: Check for the completeness constraint in CCRO by following the decomposition of a citation reason into sub-classes and partitions.</i>	
4. Scientific Knowledge Omission	
Are all the concepts in the ontology properly described?	Yes, No
If No , list the concepts that are not properly defined.	
<i>Hint: Check if CCRO has provided sufficient definition for each concept defined in ontology.</i>	
5. Redundancy of Disjoint Relations	
Is there any redundancy in defining disjointness between concepts?	Yes, No
If Yes , list the disjoint concepts that are defined more than once.	
<i>Hint: Check for the citation reason properties in CCRO where disjoint relationship is defined more than once.</i>	

Appendix E:

CCRO Package

```
\RequirePackage{ifthen}
\RequirePackage{natbib}
\RequirePackage{xcolor}

%% 'sans serif' option
\DeclareOption{sans}{
\renewcommand{\familydefault}{\sfdefault}
}

%% 'roman' option
\DeclareOption{roman}{
\renewcommand{\familydefault}{\rmdefault}
}

\definecolor{positive}{RGB}{0, 200, 0}
\definecolor{negative}{RGB}{255, 0, 0}
\definecolor{neutral}{RGB}{255, 150, 0}

\newcommand{\citepos}[2]{
\ifthenelse{\equal{#2}{+1}}
{\citep*[\footnotesize \textcolor{positive}
```

```

{\Incorporate} \normalsize] []{#1} }
{
\ifthenelse{\equal{#2}{+2}}
{\citep*[\footnotesize \textcolor{positive}
{Extend} \normalsize] []{#1} }
{
\ifthenelse{\equal{#2}{+3}}
{\citep*[\footnotesize \textcolor{positive}
{BasedOn} \normalsize] []{#1} }

{\citep*[\footnotesize \textcolor{positive}
{Incorporate} \normalsize] []{#1}}
} } }

\newcommand{\citeneg}[2]{
\ifthenelse{\equal{#2}{-1}}
{\citep*[\footnotesize \textcolor{negative}
{Contrast} \normalsize] []{#1} }
{
\ifthenelse{\equal{#2}{-2}}
{\citep*[\footnotesize \textcolor{negative}
{Criticize} \normalsize] []{#1} }
{
\ifthenelse{\equal{#2}{-3}}
{\citep*[\footnotesize \textcolor{negative}
{Negate} \normalsize] []{#1} }

{\citep*[\footnotesize \textcolor{negative}
{Contrast} \normalsize] []{#1}}
} } }

```

```
\newcommand{\citeneu}[2]{
\ifthenelse{\equal{#2}{=1}}
{\citep*[\footnotesize \textcolor{neutral}
{Discuss} \normalsize] []{#1} }
{
\ifthenelse{\equal{#2}{=2}}
{\citep*[\footnotesize \textcolor{neutral}
{Compare} \normalsize] []{#1} }
{
\citep*[\footnotesize \textcolor{neutral}
{Discuss} \normalsize] []{#1}
} } }

%% Global indentation option
\newif\if@neverindent\@neverindentfalse
\DeclareOption{neverindent}{
\@neverindenttrue
}
\ExecuteOptions{roman}
\ProcessOptions\relax

%% Traditional LaTeX or TeX follows...

\if@neverindent
\neverindent
\fi

\endinput
```


Appendix F:

Existing Ontologies

BiRO - Bibliographic Reference Ontology

The Bibliographic Reference Ontology (BiRO) is an ontology meant to define bibliographic records, bibliographic references, and their compilation into bibliographic collections and bibliographic lists, respectively.

C4O – Citation Counting and Context Characterization Ontology

The Citation Counting and Context Characterisation Ontology (C4O) is an ontology that permits the number of in-text citations of a cited source to be recorded, together with their textual citation contexts, along with the number of citations a cited entity has received globally on a particular date.

TABLE 4: BiRO, the Bibliographic Reference Ontology

No	Properties
1	is referenced by
2	references

TABLE 5: C4O, the Citation Counting and Context Characterization Ontology

No	Properties	No	Properties
1	denotes	7	pertains to
2	has context	8	has content
3	has global citation frequency	9	has global count date
4	has global count source	10	has global count value
5	is denoted by	11	has in text citation frequency
6	is relevant to		

FaBio – FRBR-aligned Bibliographic Ontology

The FRBR-aligned Bibliographic Ontology (FaBio) is an ontology for describing entities that are published or potentially publishable (e.g., journal articles, conference papers, books), and that contain or are referred to by bibliographic references.

DoCO – Document Component Ontology

The Document Components Ontology (DoCO) is an ontology that provides a structured vocabulary written of document components, both structural (e.g., block, inline, paragraph, section, chapter) and rhetorical (e.g., introduction, discussion, acknowledgements, reference list, figure, appendix).

PRO – Publishing Role Ontology

The Publishing Roles Ontology (PRO) is an ontology for the characterization of the roles of agents – people, corporate bodies and computational agents in

TABLE 6: FaBiO, the FRBR-aligned Bibliographic Ontology

No	Properties	No	Properties
1	has creator	16	has rights
2	has discipline	17	has subject term
3	has embodiment	18	is discipline of
4	has exemplar	19	is embodiment of
5	has format	20	is exemplar of
6	has language	21	is in scheme
7	has license	22	is manifestation of
8	has manifestation	23	is part of
9	has part	24	is portrayal of
10	has place of publication	25	is realization of
11	has portrayal	26	is representation of
12	has primary subject term	27	is scheme of
13	has publisher	28	is stored on
14	has realization	29	stores
15	has representation		

TABLE 7: DoCO, the Document Components Ontology

No	Properties	No	Properties
1	contains	2	is contained by

TABLE 8: PRO, the Publishing Roles Ontology

No	Properties	No	Properties
1	at time	8	is role in
2	holds role in time	9	relates to
3	is document context for	10	relates to document
4	is organization context for	11	relates to organization
5	is person context for	12	relates to person
6	is related to role in time	13	with role
7	is role held by		

the publication process. These agents can be, e.g. authors, editors, reviewers, publishers or librarians.

PSO – Publishing Status Ontology

The Publishing Status Ontology (PSO) is an ontology designed to characterise the publication status of documents at each stage of the publishing process (draft, submitted, under review, etc.).

SWAN – Discourse Ontology

Developing cures for highly complex diseases, such as neurodegenerative disorders, requires extensive interdisciplinary collaboration and exchange of biomedical information in context. Our ability to exchange such information across sub-specialties today is limited by the current scientific knowledge ecosystem’s inability to properly contextualize and integrate data and discourse in machine-interpretable form. This inherently limits the productivity of research and the progress toward cures

TABLE 9: PSO, the Publishing Status Ontology

No	Properties	No	Properties
1	at time	6	is status in
2	holds status in time	7	results in acquiring
3	is acquired as consequence of	8	results in losing
4	is lost as consequence of	9	with status
5	is status held by		

TABLE 10: SWAN, Discourse Ontology

No	Properties
1	cites As Supportive Evidence
2	research Statement Qualified As
3	refers To

for devastating diseases such as Alzheimer’s and Parkinson’s. The SWAN (Semantic Web Applications in Neuromedicine) ontology is an ontology for modeling scientific discourse and has been developed in the context of building a series of applications for biomedical researchers, as well as extensive discussions and collaborations with the larger bio-ontologies community.

CiTO – Citation Typing Ontology

The Citation Typing Ontology (CiTO) is an ontology that enables characterization of the nature or type of citations, both factually and rhetorically.

TABLE 11: CiTO, Citation Typing Ontology

No	Properties	No	Properties
1	agrees with	21	disagrees with
2	citation	22	discusses
3	cites	23	disputes
4	cites as authority	24	documents
5	cites as data source	25	extends
6	cites as evidence	26	gives background to
7	cites as metadata document	27	gives support to
8	cites as potential solution	28	likes
9	cites as recommended reading	29	parodies
10	cites as related	30	plagiarizes
11	cites as source document	31	refutes
12	cites for information	32	replies to
13	compiles	33	retracts
14	confirms	34	reviews
15	contains assertion from	35	ridicules
16	corrects	36	speculates on
17	credits	37	supports
18	critiques	38	updates
19	derides	39	uses conclusions from
20	describes	40	uses data from
		41	uses method in