

A Framework for Semantic Information Extraction from Social Media Data

**K.P.G Kodikara
2018**



A Framework for Semantic Information Extraction from Social Media Data

**A dissertation submitted for the Degree of Master of
Computer Science**

**K.P.G Kodikara
University of Colombo School of Computing
2018**



Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: K.P. G Kodikara

Registration Number:2015/MCS/039

Index Number: 15440391

Signature:

Date: 14/07/2018

This is to certify that this thesis is based on the work of

Ms. K.P.G Kodikara

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr A.R Weerasinghe

Signature:

Date:

Abstract

Information available in digital format has exceeded the capability of human brain to process, thus the processing mechanism has become an important responsibility of machines. Understating and extracting valuable information from such text have been a challenge over the past decade. Many valuable research works are done in this area and as a result there are many efficient existing approaches. Most of them are domain-based approaches and employ only shallow syntactic features in the extraction process. Also, these available approaches have limited capabilities of analysing the content semantically. This work introduces a novel approach which identifies the basic relations of the entities in a digital text in terms of triplets, in the given domain of music news retrieved from social media and extracts information to a structured format. This approach has been implemented as an independent system which receives content in unstructured format as the input and gives a structured file as the output. Proposed approach consists of three main components, namely Data Retrieval, Pre-processor, Triplet Extractor and gives a special attention to analyse content in a digital text semantically. This would be a unique solution when finding considerable amount of information from free text as it consists of a methodology of semantically analysing text content and provide requested information by the user. Question and Answering module has been implemented to perform evaluation of the main system which is included as the fourth component of the system. Through the Evaluation performed using this Q&A module, 67% of accuracy was shown, thus this research work presents a powerful generic information extraction mechanism to grab knowledge from digital text available in various resources.

Acknowledgement

This Research study was carried out with the valuable guidance provided by the Supervisor Dr. A.R. Weerasinghe.

My heart full thank goes to Supervisor Dr. A.R. Weerasinghe who always guided me right from the beginning of the project by giving valuable suggestions and necessary instructions. I would be very thankful for the valuable knowledge he shared with me in the area of Natural Language Processing, Information Extraction which was the main research area of the project.

Secondly, I would like to thank Mr. D. N. Koggalahewa- supervisor of my undergraduate research project, who introduced me to this research area and sharing his valuable knowledge and experience. Next, I would like to thank my parents, for their encouragement, guidance and support given throughout this project.

Finally, special thank goes to Mr. Gihan Seneverathne who introduced me to Dr. A.R Weerasinghe and recommending this research idea.

Thank you very much everyone, if not for you, all this project would not be a success.

Table of Contents

Abstract	iv
Acknowledgement	v
1. Introduction.....	1
1.1 Introduction to the Research Problem	2
1.2 Motivation.....	2
1.3 Objective of the Project.....	2
1.4 Scope.....	3
1.5 Summary	3
2. Background	5
2.1 Overview of Information Extraction.....	5
2.2 Literature Survey	6
2.3.1 Natural Language Generation Approach	6
2.3.2 Template based Information Extraction	8
2.3.3 Information Extraction from Social Media.....	9
2.3.4 Semantic Music Information Extraction using Rule based Approach	12
2.3.5 Extracting Users Listening Events from Social Media Data	13
2.3.6 Ontology based Information Extraction Approaches	14
2.3 Summary	15
3. Design and Methodology	17
3.1 Analysis of the Problem.....	17
3.2 High level Design.....	18
3.3 Methodology	19
3.3.1 Data collection and text-pre-processing module.....	20
3.3.2 Triplet Extractor	20
3.3.3 Post processing and Storage	22
3.3.4 Question Answering Module	23
3.4 Summary	23
4. Implementation.....	25
4.1 Introduction.....	25
4.2 Input Process and Output.....	25
4.3 Basic Components of the System	29
4.3.1 Data Retrieval Module	29
4.3.2 Filtering and Pre-Processing Module	32
4.3.3 Triplet Extraction Module	33
4.3.4 Question Answering Module.	34
4.4 Tools and Technologies Used.....	35
4.2.1 Rome API.....	35
4.2.2 Overview of Jsoup Library.....	36
4.2.3 Overview of Stanford Core NLP (Which provides the dependency parser).....	37
4.5 Summary	37
5. Evaluation and Testing	39
5.1 Introduction.....	39
5.2 Evaluation Model.....	39

5.3	Test Data for the Evaluation	41
5.4	Results and Discussion	42
5.5	Summary	43
6.	<i>Conclusion and Future Work</i>	45
6.1	Introduction.....	45
6.2	Research Findings	45
6.3	Problems Occurred	46
6.4	Limitations.....	46
6.5	Future Work	47
	<i>References</i>	49

List of Figures

Figure 2-1	7
Figure 2-2	9
Figure 2-3	11
Figure 3-1	19
Figure 3-2	20
Figure 3-3	21
Figure 3-4	21
Figure 4-1	26
Figure 4-2	26
Figure 4-3	26
Figure 4-4	28
Figure 4-5	28
Figure 4-6	29
Figure 4-7	29
Figure 4-8	30
Figure 4-9	30
Figure 4-10	31
Figure 4-11	31
Figure 4-12	32
Figure 4-13	32
Figure 4-14	33
Figure 4-15	34
Figure 4-16	35
Figure 4-17	35
Figure 5-1	41
Figure 5-2	42
Figure 5-3	43

Abbreviations

NLP- Natural Language Processing

IE- Information Extraction

IR- Information Retrieval

POS Tagging- Part-Of-Speech Tagging

NER-Named Entity Recognition

RSS- Really Simple Syndication

Q & A Module- Question Answering Module

WELOTA- Well Loved Tales

1. Introduction

In the recent couple of decades, a rapid growth of freely available textual information (in digital format) has occurred in the internet and other resources. A substantial amount of such free textual data related to medical, legal and especially in social media, communication and many more fields are available in many resources. Those information is used in various studies to build powerful and efficient information extraction tools. Information Extraction(IE) is a significant sub task of Natural Language Processing. It deals with finding factual information from unstructured text. Facts capture real world entities, events occurrences or states with attributes, actors or arguments. In more precise terms information extraction is done to identify relevant parts of text usually which belonged to a predefined specific domain by ignoring other irrelevant information. As a result of the IE tasks a structured representation of relevant information is produced typically in relations or in a knowledge base. Throughout the history of NLP, Information Extraction has always remained a challenge. A prominent reason for this is that Natural Language content available in digital text is not directly machine processable. Information Extraction is often referred to as automatic extraction of structured information. That structured information consists of elements like entities, attributes, relationships between entities and attributes which describes entities contained in unstructured or semi-structured sources.

Through information extraction tasks, it is expected to organize the information belonging to various domains and to put them in a semantically precise format that allows further inferences to be made by computer algorithms. Extraction process involves, identification of structured elements like noun phrases, person, locations, numerical expressions and finding semantic relationships between those entities.

Though during the couple of decades IE has been a continuous interested area, it has been a really tough task even in a very limited domain. Causes for this can be highlighted as the ambiguity and the complexity of natural language. Many powerful and efficient tools have been developed as a result of valuable research work. Most of them are domain specific approaches and employ only shallow syntactic features in the extraction process. Also, these available approaches have limited capabilities of analyzing the content semantically.

1.1 Introduction to the Research Problem

There is a vast amount of textual data (unstructured digital data) available in online, web documents, blogs, social media, online news etc. From those various resources, highest amount of data is available in social media. Due to the higher usage of social media, Information belongs to other resources also have included into social media pages. Processing all those massive amounts of data available has exceeded the capability of human brain to process hence the processing mechanism has become an important responsibility for machines. Important and meaningful data can be extracted by processing unstructured data. When capturing those information in the basic level, it retrieves more generic information about natural language. For instance, most of IE tools primarily describe syntactic information (POS tagging, chunking, parsing etc.), semantic role labelling or named entity recognition. Most of those available tools employ only shallow syntactic features and contains limited capability of analysing textual content semantically. They are capable of only annotating data in to a single semantic class such as person, location, numerical expressions, currency, day etc. When extracting meaningful content from textual resources it has been a demanding necessity for retrieving additional information. Hence extracting information from textual data involves deep understanding of natural language by machines.

In this research work, identified research problem can be stated as below,

- There is no well-established mechanism to Extract Information Semantically from Raw Textual Data and Discovering Valuable and Relevant Information from it.

1.2 Motivation

This research idea is inspired through many previous works. But mainly this is evolved from “An Ontology Based Natural Language Story Generation Approach” project. This research idea is based on the Knowledge Extraction Module of the above research work and enhancement of the module will be carried out through this project [1].

1.3 Objective of the Project

The Main Objective of this Research Project is to explore and identify a mechanism to build an efficient and powerful information extraction algorithm, which analyse the information semantically in a digital text and map the content into a meaningful, structured format. Above main objective can be broken down to the below stated sub objectives.

- To continue the literature survey on similar existing Information Extraction systems and identifying the basic limitations of them.
- To select sample data from the selected area (Social Media/RSS Feeds) and to build a mechanism to filter only the domain specific (News Items of Music Artists) information.
- To build an efficient algorithm to extract News Items of Music Artists from the unstructured text and to annotate them using the semantic analysis approach.
- To build an evaluation method to evaluate the extracted information and perform a statistical evaluation on the success rate of the proposed information extraction algorithm.

1.4 Scope

For the development of this project, “News Items of Music Artists and their Tracks” is selected as the domain of the system. Extraction of information from social media is considered to be more challenging task than classical Information Extraction ie. extracting information from trusted resources. As it is highly challenging task to extract valuable content from very short, noisy, misspelled content, in order to introduce a generic approach for semantic information extraction, News Items of Music Artists from social media is taken as a sample case study for the development of this research project. Extracting semantic information and relations about News Items of Music Artists and their Tracks, from social media will be covered through this project.

- Only Social Media Data in English Language will be considered.
- When extracting the meaning of the content only the domain of Recent News Items of Music Artists will be considered.

1.5 Summary

Objective of this research work is to overcome the above-mentioned research problem by using Music Artist News as the base domain. Those data will be retrieved from social media/RSS Feed as large amount of information is freely available and they provide information that is more up to date and convenient to access.

Next chapter provides a detailed description of the background related with the information extraction and discuss on the previous work that has been done in the research area of Information Extraction.

2. Background

Previous chapter provided some basic idea through the introduction to this project. This chapter focuses on the knowledge gathered through a thorough literature survey conducted during the initial stage of the project.

Designed approach of the project can provide a powerful general-purpose framework for extracting information from unstructured digital text and providing useful information which can be easily represented in suitable systems (can be given as an input to a knowledge representation module that can efficiently retrieve the information given). Before moving on to the designing of the novel approach, general idea on information extraction system is presented.

2.1 Overview of Information Extraction

Information Extraction can be introduced as the process of automatically identifying a set of predefined concepts from unstructured or semi structured set of information which belongs to a specific domain. Typically, the output of the Information Extraction process will be in structured format which contains the information in a machine-readable format. Information extraction is an important task in text mining. It is formally stated as task of finding structured information from unstructured or semi-structured text [2]. Main goal of Information Extraction is to identify important information from unstructured or semi structured text based on a specific domain. An ideal example for Information Extraction is given in [2] ,

“In 1998, Larry Page and Sergey Brin founded Google Inc.”

Following information can be extracted from the above English sentence.

- a) FounderOf(Larry Page, Google Inc.),
- b) FounderOf(Sergey Brin, Google Inc.),
- c) FoundedIn(Google Inc., 1998).

Structured, Unstructured and Semi Structured documents

Structured Documents- These types of sources represent the data in a structured meaningful manner that the knowledge contain in the source is directly visible to the machine or human.

Unstructured Documents-These source cannot be process by machines directly. This contains information in unstructured and unorganized manner, that the humans or machines required to put extra effort to retrieve meaningful information from those content.

Semi Structured Documents- This type of sources falls in to the structured category but inadequate of presenting information in a complete structured or organized manner. It requires some amount of effort to grab meaningful content from these types of sources.

When performing Information Extraction Tasks, other irrelevant information is ignored, and it is required to retrieve only important information from the unstructured text. The process of extracting information involves identification of simple format of structures such as noun phrases and one or many semantic relations between them. In the relation a) of the above example noun phrases can be listed as Larry Page and Google Inc. Semantic Relation between the two noun phrases is the term “FounderOf”. From those two noun phrases, first and second noun phrases can be identified as subject and the object.

2.2 Literature Survey

2.3.1 Natural Language Generation Approach

As stated in the section “Motivation” for the research work, apart from the previous research work done in[1] there were many research work that inspired this proposal. Preliminary the research work of [1] will be described.

WELOTA presents a powerful mechanism to extract information from unstructured textual data, represent the extracted information using a knowledge representation mechanism and retrieve the knowledge to construct meaningful natural language sentences. To prove this approach Well Loved Story (Fairy Tales) domain is selected. This research work consists of three modules - Knowledge Extractor, Knowledge Representor and Natural Language Generator. Main components of the system can be illustrated as described in Figure 2-1.

Knowledge Extractor is responsible for the information extraction. It takes unstructured text file as an input. This file contains story in the most primitive form in simple English. This file is processed and annotated according to the defined rules and a structured file is generated. This structured XML is the input to the Knowledge Representor which generates the OWL file automatically. Natural Language Generator is responsible for generating natural language text. Knowledge Inferencer takes the title of the story as input and retrieve data from the ontology and passes them as intermediate data structures to story generator. The Story Generator then generates the final output as natural language sentences. Proposed research work is interested in the enhancement of the Knowledge Extraction module of WELOTA. Knowledge extractor module has used the ANNIE pipeline of the GATE(General Architecture for Text Engineering)[3] for basic natural language processing tasks.

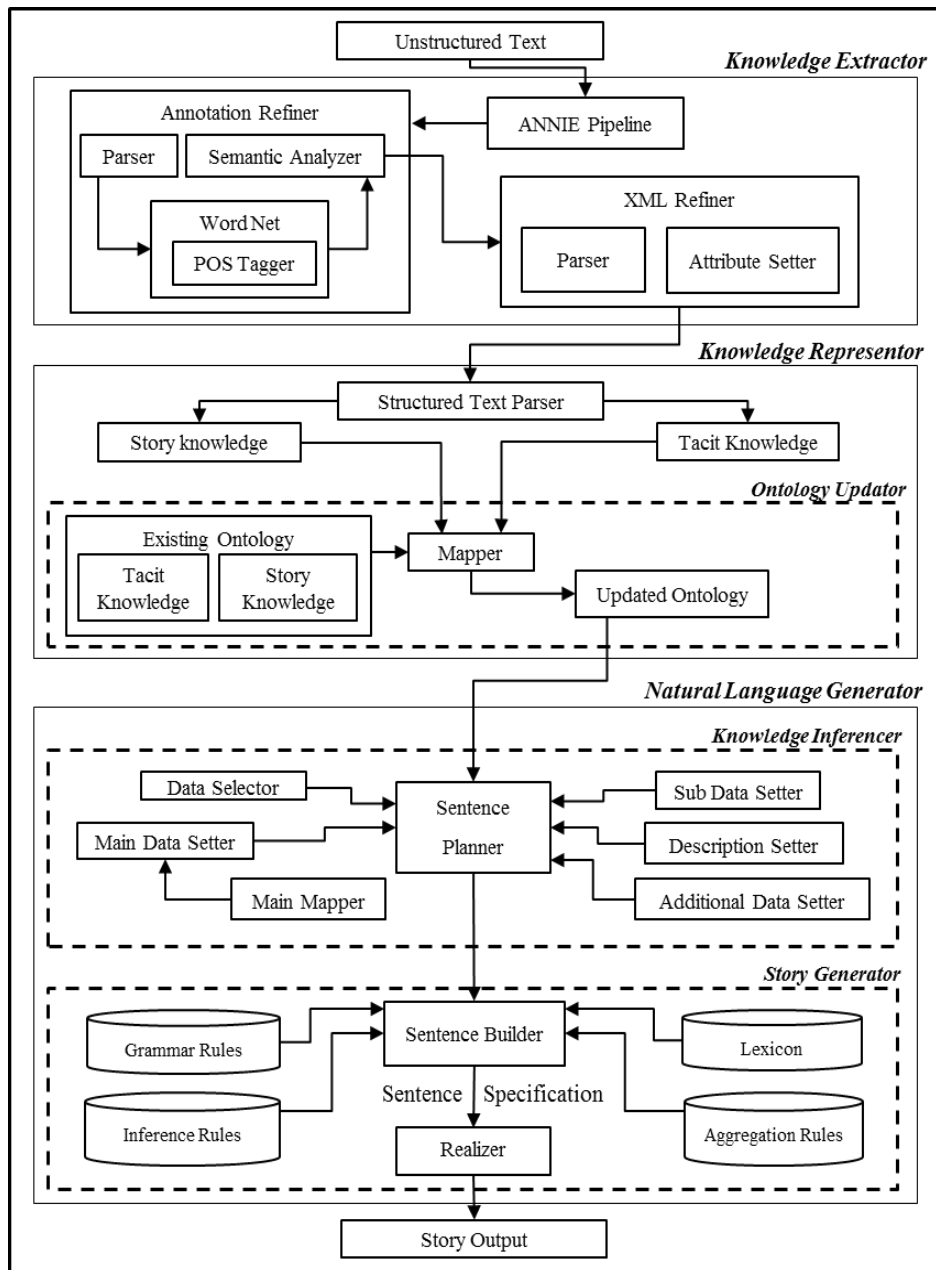


Figure 2-1

Namely those tasks are Sentence Segmentation, Tokenization and Named Entity Detection of basic parts (Such as identifying story characters, locations, country, rivers, quoted text, numbers etc.). During this NE task the system is only capable of mapping a single entity into a one semantic class. Then the received output is passed through two components, POS Tagger and Semantic Analyser which does the POS Tagging and convert the simply labelled entities into multiple semantic classes which is organized hierarchically. To implement this component RiTa WordNet[4] libraries are used. Example output is shown below,
 <Object description=" red"><Food><Fruit><Sweet>Apple</Sweet></Fruit></Food> </Object>.

Though this approach presents a very powerful mechanism of extracting information by mapping the entities into multiple semantic classes and retrieve information by analysing content semantically considering as a generic approach, it lacks some important features which will be not extracted. When extracting information by analysing content semantically, it only considers a specific set of contexts. Also, this approach provides only relations between entities within a sentence. If an entity occurs more than once, it does not consider it as an existing entity and it extracts the information from the scratch. This approach does not try to map the relation between similar entities which does not belong to the same sentence. Finally, it can be concluded that this work is only able to partially address the research problem.

2.3.2 Template based Information Extraction

Apart from the above research work, project Artequakt [5] was one of the main motivation to originate the research idea. Artequakt is a system which extracts knowledge about artists from the web, represents them in ontology and uses predefined templates to generate personalized biographies of the artists based on its ontology. It has a Knowledge extraction component which focuses on extracting relevant information of an artist from several web pages. So In order to gather relevant information from different places they have mainly focused on identifying relationships between entities. As an example, they have mentioned “Rembrandt” is a person or “15 July 1606” is a date. Typical Information Extraction System are capable of mapping those entities. But the challenging task is to identify the relationship between those two entities is identified as “Rembrandt was born on 15 July 1606”. This is identified using an ontology coupled with WordNet a general purpose lexical database[6], and GATE[3]. So in order to identify relationships first they syntactically analyses the extracted phrases and then the semantic analysis is done to identify relationships between them. In order to detect relationships first they identify the main components of a sentence using a semantic analysis. This semantic analysis focuses on identifying only the basic entities like subject, verb, object etc. In other words it considers only mapping a single entity into a one semantic class. It does not focus on analysing the meaning up to a deep level. It just highlights only the narrow meaning which assists only to identify relationship between two entities.

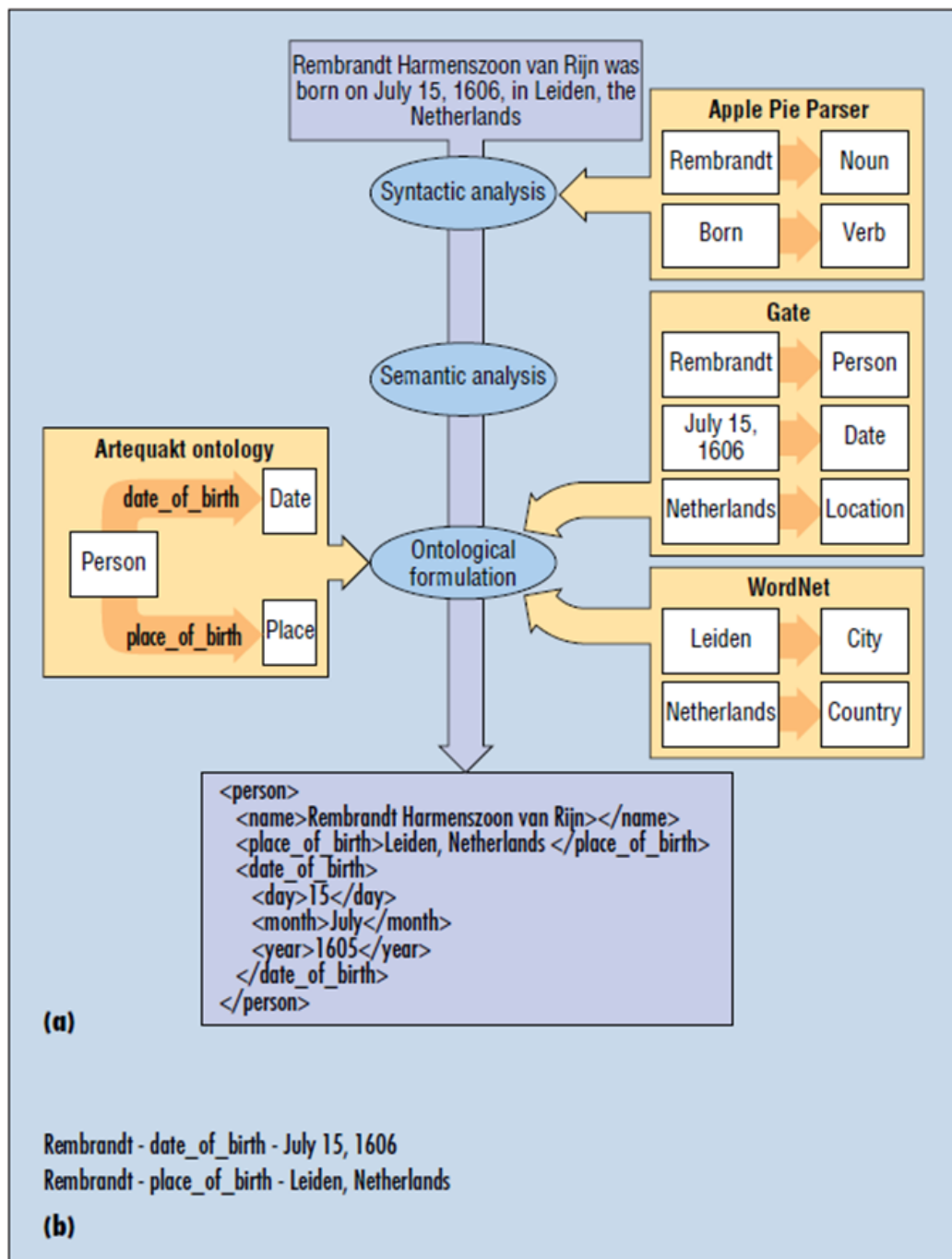


Figure 2-2

This research work successfully identifies the relationship between entities but does not map those entities to multiple semantic classes. Figure

2-2 shows the architectural diagram of the Knowledge Extraction Module.

2.3.3 Information Extraction from Social Media

Information Extraction from Social Media[7] was selected to be studied related with this research work, as it introduces a framework for Information Extraction from user generated unstructured content on social media. This gives an overall idea of an approach to Extract

Information from social media. In the introduced work they have provided an open source set of portable and customizable modules that can be used by different application belong to various domains as Financial, Security, and Web Search etc. Scope covered by the work belongs only to the social media. They have proposed this approach also by addressing common challenges of Information Extraction from social media content. It contains four main modules Noisy Text Filtering, Named Entity Extraction, Named Entity Disambiguation and Fact Extraction. Noisy Text Filtering component filter the non- informative posts/content in posts. This filtering is done using domain knowledge and criteria that only keeps the relevant domain related information/ posts. Named Entity Extraction component, they have come up with a novel approach which does not contain identifying heavy syntactic features and gives clues derived from the Named Entity Disambiguation model. In the Named Entity Disambiguation model, they have proposed an open world approach. Named Entities are disambiguated by linking those to their original source (they have used primary source as Wikipedia and secondary sources as a social network page). In the introduced approach they have used a feedback loop approach which is unique feature of the proposed research. Each component/subtask of the framework provide a feedback to the previous subtask which allows the opportunity to modify or refine the output of one subtask. It provides a brilliant example of this task. When NED module trying to disambiguate the mention “Apple”, it finds out that it cannot be linked to an entity. Can derive the conclusion as this mention “Apple” refers to the fruit rather than the company. If using the traditional approach once the NED has identified the mention cannot be identified as an entity it is eliminated. Figure 2-3 is the diagram of the proposed IE framework of this work along with the traditional approach. Although this approach presents a unique approach by introducing a framework which eliminates the common problems faced by many previous work, it has not fully achieved the research problem of the proposed research idea- which is, providing well-established approach to Extract Information Semantically from Textual Data.

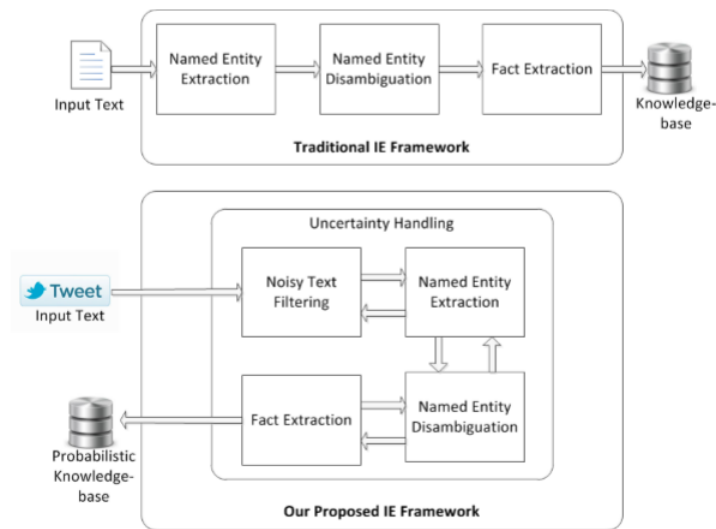


Figure 2-3

Through Feedback loop approach they have provided a well-established mechanism to refine the Named Entity Disambiguation and provide more meaningful content. If the approach proposed by this work i.e. multiple semantic role labelling, can derive a very successful framework for Semantic Information Extraction from social media content.

TWICAL[8] provides an open domain calendar of significant events from Twitter data. They have introduced a successful approach for extracting important event categories from twitter and classifying extracted events based on latent variable models. This approach contains four basic components for the below tasks,

- 1) Named entity segmentation
- 2) Extracting event mention
- 3) Extracting And Resolving Temporal expressions
- 4) Classification of event types

TWICAL[8] extracts 4 tuple representation of events including named entity, event phrase, calendar dates and event types. When a stream of tweets are provided to the system, it first extract Named Entities(NE) through the Named entity segmentation by associating with event phrases and dates (unambiguous) related to important events. In the Named Entity Segmentation task, to avoid the limitations and performance issues of the existing NLP tools due to the noisy data retrieved from twitter, they have built a separate named entity tagger trained in domain of Twitter data. In the Event mention Extraction component, uses the annotated tweets corpuses to train sequence models to extract events. An important context is provided by those event phrases. A good example given is for extracting the entity Steve Jobs. Entity Steve Jobs and the event phrase died in connection with October 5th, is much more

informative than simply extracting Steve Jobs. When viewing this approach in the semantic information extraction point of view, they have been successfully achieved the entity relationship than the WELOTA, Artequakt and the novel framework for IE. This approach can be included to the currently proposed solution for identifying relations of entities. Extracting And Resolving Temporal expressions and Classification of event type components are responsible for resolving temporal expressions (eg. Next Friday, Yesterday, tomorrow etc.) for exactly extracting the exact calendar references. To resolve Temporal Expression they have used an existing tool called TempEx[9]. To classify the event types they have used unsupervised approaches that will automatically induce event types which match the data. They have adopt their work based on latent variable models inspired by previous work.

2.3.4 Semantic Music Information Extraction using Rule based Approach

Towards Semantic Music Information Extraction from the Web Using Rule Patterns and Supervised Learning[10] presents an semantic information music information extraction approach to develop automatic methods to discover semantic relations between musical entities from Web. This approach contains two main subtasks, automatic band member detection and automatic discography extraction. That is determining which persons a band consists (or consisted) of and recognition of released records (i.e., albums, EPs, and singles).

To extract semantic Music information from text, two strategies manual rule-pattern approach and the supervised learning approach. Before performing those strategies, typical pre-processing tasks (Like removing mark-ups, tokenization, sentence segmentation, POS Tagging, Gazetteer Annotation, Transducing) were performed to the textual data retrieved from Web. In the Rule pattern approach consist of set of Rules which is built by capturing music related information on Last.fm[11], Wikipedia and allmusic[12] websites. Those specific rules has implemented using Jape grammars that are used in the transducer step of GATE. In the Supervised Learning approach, instead of manually analysing contents, they provide an approach to apply supervised learning algorithm to set a pre-annotated examples. They have used a learning model to find the relevant information regarding artists and their tracks in web documents. Several types of machine learning algorithms have introduced for automatic information extraction tasks. After the extraction step using the rule or learning based approaches, for each processed text and interest, a list of potential entities are acquired. For each band/ artist or songs all the related text with each band or song are joined and the occurrences of each entity and the occurrence frequency is counted. This process is performed

as a part of Entity Consolidation and Prediction. This work seems to be directly related with the proposed approach and have achieved the intention of the research idea for a certain extent by using a semantic information extraction approach. But the approach used in this related work seems to be limited to meet the requirements of the proposed idea as, similar to the other IE tools studied in this work, lack of analysing the meaning of the content into a deep level.

2.3.5 Extracting Users Listening Events from Social Media Data

Another research work highly related with the domain of the proposed research is NowPlaying [13]project. This present a set of nowplaying data set which features listening events of users which includes listening patterns of users to tracks and artists over two years. They have used twitter to extract information of tweets that describes a user has listened to a particular artists' track (analyse #nowplaying tweets). In this approach they have provided a publicly available and extensive dataset of listening events which is updated daily and gathered from the Twitter platform and interlinks their dataset with the MusicBrainz database (allows for a central reference point for artist and track information which can be used to further gather information from other datasets and source.) For the creation of the #nowplaying data set they have presented an Extraction Framework which consist of four components. Those components are Twitter API Crawler, Basic Extractor, Track and Artist Extractor and Spotify Extractor. The basic extractor is responsible for extracting simple metadata from the data gathered via the Twitter API, Track and Artist Extractor relies on the content of the tweets and aims at extracting the artist and track mentioned in the tweet. Tweets were not able to directly resolve against MusicBrainz were sent via the Spotify music streaming platform and employ the Spotify extractor which exploits information from the Spotify website [14].

From the basic components we are interested in Track and Artist Extractor and Spotify Extractor.

From the basic extractor the date and time when the given tweet was sent, the service used for publishing the tweet and the username: the user name of the user who sent the tweet is directly extracted. In the Track and Artist Extractor matches artists and songs that are included within tweets with the relevant entries in the MusicBrainz database. In the final dataset, they have only included listening events which were able to resolve against the MusicBrainz database. Also they have provided data which can be interlinked by using e.g., the MusicBrainz identifiers for artist and/or song. Example of the output of this module can be shown as below,

- Input tweet: "Like a Rolling Stone - Bob Dylan #nowplaying #listenlive".
- Output: song ("Like a Rolling Stone") and artist ("Bob Dylan").

Steps of the extraction process can be listed as below,

- 1) Clean the tweet text by removing URLs and whitespaces.
- 2) Split the text of the cleaned input tweet using delimiters (; - etc.).
- 3) Check artist contained in the MusicBrainz database is contained in the currently checked text chunk.
- 4) If an appropriate artist can be found, retrieve all songs performed by these artists from MusicBrainz and check, whether one of the song titles is contained in the tweet.
- 5) Check for duplicates.

Spotify Extractor is facilitated to increase the quality and quantity of listening events in the #nowplaying dataset is the Spotify Extractor, which leverages tweets which were sent via the Spotify platform. This also contains a similar approach to the MusicBrainz Extractor.

Like all the other related research works, this also lacks prominent feature we expect from presenting this research idea. Provided work only extracts the music and artist data and relate them together. But it does not provide any extra information about the artist/ track type. By combining the proposed research idea, we would be able to extract not only the most popular tracks or artist, we could retrieve other information as the genre of a specific track and which kind of music is popular among certain period of time, find out the genre a particular artist is popular on etc.

Apart from the above related work some other work was also examined to identify the approaches that common Information Extraction tools/IE module of research work has done.

2.3.6 Ontology based Information Extraction Approaches

Ontology guided Information Extraction from Unstructured Text[15] is an approach which describes to populate an existing ontology with instance information present in a natural language text. In the Knowledge Extraction process of this approach, first they identify the domain with the aid of a domain inference module which incorporates a Semantic Lexicon predefined by experts. Then using instance extractor module they extracts instance information. Here they basically identify the subject, predicate and the object of a particular sentence. Using NLP they identify the subject and map in to the subject into the semantic class, and use the predicate and object as the attribute name and value respectively. This is the basic task that is handled by this approach when mapping structured file into the OWL file. In this approach as

the methodology of previous research work mentioned, only basic raw text extraction is handled. Further they have not mentioned on semantic IE in this approach.

Text-TO_ONTO[16] is an ontology learning system which is based on a general architecture which discovers conceptual structures and engineering ontologies from text. In the Text Processing Server it uses a shallow text processor based on the core system SMES(Saarbrücken Message Extraction System)[17] to extract knowledge in the text. It uses a message Extraction system which performs syntactic analysis on text documents. Further it has used a lexical resource in order to collaborate semantic information with the syntactic information. The extracted information is mapped into a XML file. But it has not mentioned of a broad level of semantic analysis in its text processing server.

Text2Onto[18] which is similar system to Text-TO_ONTO a framework for ontology learning from textual resources. To analyse text it uses basic linguistic processing combined with machine learning techniques. For basic text processing tasks it uses GATE framework[3]. After Linguistic pre-processing is handled, it uses JAPE transducer to identify the patterns required to the ontology learning algorithm. They have created JAPE patterns for shallow parsing and to identify instances, concepts and different types of relationships. Apart from the above specified basic linguistic tasks and relationship identification tasks further they have not mentioned any semantic specific analysis.

2.3 Summary

Many Researches have worked on approaches which has been a large contribution to this Research Area. In this literature review only a few such approaches have been discussed which has been an immense motivation to provide the research outcome of the proposed work. Most of the above discussed solutions contains a different approach and many consist of ontology-based approaches, rule based or lexical resource-based approaches or Supervised Learning approaches. Research work done in the Music Domain combining social media, mostly has covered Listening behaviours of fans and Extracting Artist and Track information by analysing the content on social media. Most research work done using the semantic analysis approach, are more complex, time consuming and requires lots of training and time to design. This work introduces a novel approach which addresses the research problem “There are no well-established mechanism to Extract Information Semantically from Raw Textual Data and Discovering Valuable and Relevant Information from it.”

This research work introduces an approach which is capable of extracting meaningful content by identifying relations from plain text. These relations are arranged in the format of Triplets. Triplets are in the form of subject, predicate and object. So, given a sentence, core meaning of that sentence is extracted, hence the most important knowledge of the sentence is captured through these Triplets.

Through this literature survey, general limitations were identified, and those limitations have directed to the implementation of a novel approach which tries to overcome those generic limitations, thus the designed approach is described in detailed in next two chapters.

3. Design and Methodology

Previous chapter described about the necessary background information along with a thorough literature survey. This chapter analyses the approaches that has been reviewed through the literature survey and designed a methodology according to the nature of the test data set. But this approach is not a domain specific approach as it is capable of extracting useful information of any type of data by using the semantic analysis approach. Only the question answering module is designed based on the domain -Music Celebrity News. And it also can be used to query data with slight modifications. Following is a basic overview of the analysis design and the methodology of the proposed research idea.

3.1 Analysis of the Problem

There is an ever-growing large amount of information is being available in digital format in various repositories in internet and intranets. Considerable amount of those digital information is transferred via websites, social media and various resources in unstructured format. Those substantial amount of information has surpassed the processing capability of human brain. Hence the processing huge amount of information has become an important responsibility of machines. In order to understand, the content in unstructured text the machines should have a deep understanding on natural language. Many successful approaches have been introduced to extract information from those massive amounts of freely available data throughout the previous decade. But many of those approaches are domain specific and employ only basic syntactic features and contain limited capability of analysing content semantically. That is some of these approaches are only capable of identifying the meaning of each syntactically identified content without specifying the relations between each entity. Above gaps identified, has been discussed in the previous chapter through comparing and contrasting some of related previous research work. Analysing previous works and recognising common limitations lead to identify the main research problem of the proposed research work.

Research problem: “There are no well-established mechanism to Extract Information Semantically from Raw Textual Data and Discovering Valuable and Relevant Information from it.”

Through analysis of the research problem following objectives and sub objectives were identified,

Main Objective:

- To explore and identify a mechanism to build an efficient and powerful information extraction algorithm, which analyse the information semantically in a digital text according to the identified context/domain and map the content into a machine readable, meaningful, structured format.

In order to achieve the main objective, following sub objectives were identified by analysing the main objective and breaking it down to sub parts.

Sub Objectives:

- To select a suitable domain to address the research problem.
- To build a mechanism to filter only the domain specific information.
- To build an efficient algorithm to extract factual data from the unstructured text data based on the requirement and annotate them using the semantic analysis approach.
- To build an evaluation method to evaluate the extracted information and perform a statistical evaluation on the success rate of the proposed information extraction algorithm.

After identifying and analysing those objectives a set of requirements were defined.

- System should be able to filter the data given as the input as unstructured format to related to the specific domain.
- System should perform the required text pre-processing tasks. (Cleaning and Filtering html tags, scripts etc.)
- System should perform required NLP Tasks on the output of the above processes before passing in to the algorithm which analyse the content semantically.
- Based on the predefined domain, system should perform a semantic analysis on the text by identifying the relationship between those entities, in triplet format.

3.2 High level Design

This section of the chapter describes about the proposed design which is presented as a solution to the main research problem. It describes each component of the above high-level design. To achieve the above-mentioned objective, defined sub objectives should be achieved in the given order. To easily achieve those sub objectives, requirements that has been defined accordingly should be satisfied. To achieve the above requirements a suitable design has been planned and executed according to the thorough analysis of the research problem. As a high-level solution to the above set of requirements the following design is proposed. First Data in the domain of Celebrity News is collected from RSS Feeds and cleaned and filtered in a format that could be

further processed by language parsers and the developed algorithm. Then the cleaned and filtered data is being passed to the Natural Language Processing Parser tool and the developed algorithm to extract meaningful information. These meaningful information is collected through a set of algorithms which forms the data in the format of triplets. These triplets are identified in the subject, relation and predicate format. After extracting triplets from the data, they are arranged in a structured format by further processing it. Figure 3-1 shows the high levels design diagram for the implemented system.

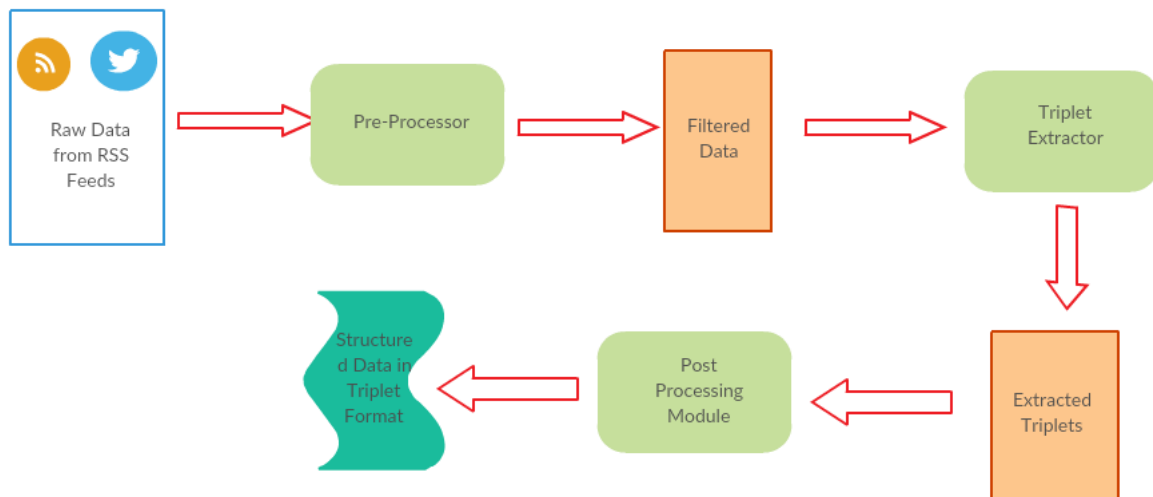


Figure 3-1

3.3 Methodology

This sub section of the Design and Methodology chapter describes about the methodology of the system. It describes how each and every component has been broken down in detail and further describes about the process of each component in detail. Process of the system starts with streaming the input for the system from the RSS Feed API. Initial input is later cleaned and filtered through the data collection and pre-processing module. Then the cleaned data is passed to the Triplet extractor module to identify the basic relations the content. This step is the most important step in this research, as this process itself addresses an enormous percentage of the research problem. Then the extracted triplets are saved on to the storage device in text format. These triplets are later queried by an Q&A (Question Answering Module) for the evaluation as well as to perform the queries intended by the user. Below sections presents a detailed description of the stages covers by the methodology of the work.

3.3.1 Data collection and text-pre-processing module

Task of Data collection and text-pre-processing is to prepare all the collected data (RSS Feeds) in natural language text format so that the cleaned and filtered information is ready and processable by the next module. Tasks of this module involves, removing HTML tags, scripts and other noisy data. After removing noisy data, output data is sent to the coreference resolution module as coreference contains vast amount of knowledge in the original content.

Coreference resolution is the task of finding all expressions that refer to the same entity in a text. It is an important step for a lot of higher level NLP tasks that involve natural language understanding such as document summarization, question answering, and information extraction [19]. Figure 3-2 shows an example for the coreference resolution. It identifies all the original entities as it is.

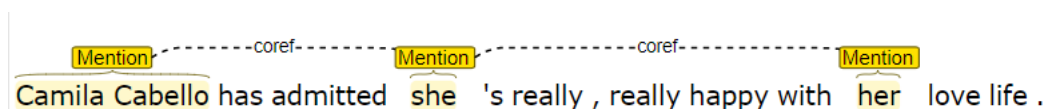


Figure 3-2

When the coreference resolution is performed, most of the coreference are mapped into their real entities. To improve the accuracy of the extraction process coreference resolution should be perform as a part of the pre-processing task before passing the data in to the triplet extraction module.

Another pre-processing task completed before triplet extraction is stop word removal. This is not an essential task as the triplet extraction module only filter only the important content in data. After this, an important pre-processing task performed is stemming and lemmatization. This is typically performed before or parallely with the triplet identification process. Output of this component would be a document consist of natural language text.

3.3.2 Triplet Extractor

This is the main component of the methodology, which is responsible for identifying the meaningful content from the input data. Cleaned and filtered textual content is provided to this component. In order to extract meaningful data, this module maps the given input data into triplet format. To convert the data into the triplet format is done using the dependency parser of the Stanford CoreNLP. Triplets are extracted in the format of (Subject,Predicate,Object). When a sentence is given to the triplet extractor, goal was to identify the main subject predicate and the object of the sentence. These subject relation/predicate and object can be identified by

the dependency parser. Dependency parser analyses the grammatical structure of a sentence, establishing relationships between "head" words and words which modify those heads [20]. This was designed to provide a simple description of the grammatical relationships in a sentence without the assistance of a linguistic expert. According to Daniel Jurafsky & James H. Martin dependency based approach can be used to provide an approximation to the semantic relationship between predicated and their arguments [21]. It further describes how it can be used for many applications including information extraction and question answering systems. In this research work Dependency parser of the Stanford core NLP module is used. Typically, a basic dependency parser forms an output as the following structure shown in the below Figure 3-3. (NNP- Proper noun, singular, VBZ- Verb, 3rd person singular present, CD-Cardinal number, IN- Preposition or subordinating conjunction, NN- Noun, singular or mass) [22]. List of basic dependencies that is usually used are: root, det, dobj, nsubj, compound, pobj, prep.

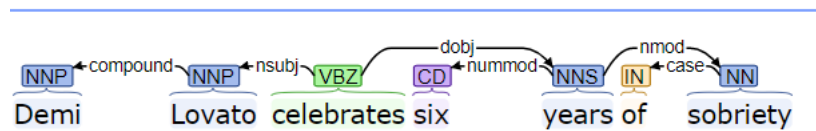


Figure 3-3

Following diagram in Figure 3-4 shows the dependency tree for the sentence “Demi Lovato celebrates six years of sobriety.”

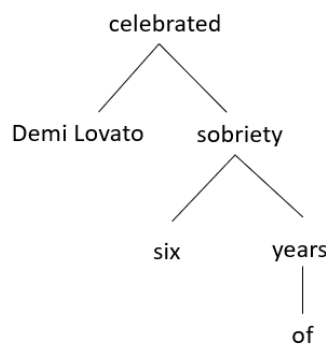


Figure 3-4

These dependencies should be converted into triplet format, in order to use the semantic approach to extract information as stated in the objectives on the Introduction chapter. To convert them to triplets there should be a set of rules which should be used to identify each of these subject, object, relation separately. When using those kind of categorization algorithms, the accuracy of the output is a concern, so there should be a well-established method to identify those 3 types of relations from a particular sentence. There are several approaches followed by

the previous researchers such as Machine Learning approaches, Tree bank parsers or rule-based approaches which are integrated with tree bank parsers. In this approach rather using an rule based or machine learning approach, Stanford OpenIE is directly used [23]. When selecting this methodology (approach), sample data set is thoroughly analysed and considered as one of the main factors. The other approaches were eliminated as most of the sentences contain in the data set are very much complex sentences which contains ten to twenty words and building and training such a model with such complex sentence will be a very expensive task to achieve. It is an obvious fact that only rule-based approach will not be an acceptable method as it is in processing these types of sentence. Using the predefined algorithm provided by Stanford OpenIE, can extract structured relation triples from text without specifying the schema for these relations in advance. A famous example that is given is “Barack Obama was born in Hawaii” would create a triple (Barack Obama; was born in; Hawaii), corresponding to the open domain relation was-born-in(Barack-Obama, Hawaii) [23]. Likewise, from any complex sentence only information related with the basic relations can be extracted. Here is an example how basic level of information is extracted from a somewhat complex sentence. Sentence: “Fifth Harmony is due to perform at the Hard Rock Event Center, in Hollywood, Florida on 11 May, in San Juan, Puerto Rico on 13 May, and in Reykjavik, Iceland on 16 May.”-Triplets: (fifth harmony; perform at; Hard Rock Event Center on 11 May in San Juan in Reykjavik on 16 May). Thought the triplets extracted from complex sentences are not straightforward as a simple sentence, the approach is capable of retrieving the core meaning of the sentence. To identify relation triplets in the OpenIE module dependency parsing is used along with several other basic annotating tasks. These annotation tasks involve, tokenizing, splitting, identifying part of speech (POS), lemmatisation, identifying full entities (NER) of each sentence. A detailed description of this module and how it is implemented is covered under the implementation section.

3.3.3 Post processing and Storage

After extracting the basic triplets, all the identified triplets are written on to local txt file which is stored as a single file, by categorizing by its news article heading. Finally, all the identified triplets are merged in to a single storage location in order to be easily processed by the Question Answering module which is used for the evaluation and to present the final out put to the end user. For the moment duplicate triplets are not considered as a limitation as such occurrences are very rare, as one sentence in a news article can be presented in multiple ways, including additional information which was not available in a previous sentence. Identifying those

duplicates would be an opening for another research problem thus it is not considered in this work.

3.3.4 Question Answering Module

This module is not related with any of the objectives of the system and it is added as an external component to the main system. This is the final component of the system and it is used as a demonstration of one of the popular applications of the semantic information extraction system. Q&A (Question Answering) Module is basically responsible for querying and evaluating the extracted knowledge. Querying process is performed with respect to a predefined domain - "News Items related to Music Artists". When designing this module, most frequent set of questions that can be generated from the given data set is analysed and categorized in to set of fragments. Then the questions are ideally converted in to the triplet format using the Stanford OpenIE. This module is developed to query two main types of questions. Querying the subject and Querying the object (Questions based on Subjects and Objects). First the Question is converted into simple sentence format, then the missing part of the sentence is identified and replaced with a constant. Then the extracted knowledge from dataset is compared against this converted triplet. Answer for the question is generated based on the results of this comparison. Detailed descriptions of this module are discussed in the implementation and the evaluation chapters. This module is separated from the main components (pre-processor and the triplet extractor) which is described in the previous sections of the methodology. Tasks of this module is performed by the end user who uses this application and for the evaluation purpose of the application. This component only considers simple scenarios when answering to the queries, as this module is designed only for the demonstration purpose and for the evaluation purposes. It rejects all the other complex form of queries.

3.4 Summary

This chapter provided a detailed description of the Design and the Methodology of the system. Most of the previous work which is done in this area are domain specific approaches and complex approaches which consist of ontology-based approaches, rule based or lexical resource-based approaches or Supervised Learning approaches. Research work done in the Music Domain combining social media, mostly has covered Listening behaviours of fans and Extracting Artist and Track information by analysing the content on social media. Most research work done using the semantic analysis approach, are more complex, time consuming and requires lots of training and time to design. Introduced approach in the above is not a domain specific approach as it is capable of extracting useful information of any type of data by using

the semantic analysis approach. This section presented how the high-level design has been formed to achieve all the objectives mentioned in chapter one. First the data is being collected using an API and those collected data is filtered and cleaned to be able to be processed by the latter components. Then the filtered data is pre-processed to retrieve a high accuracy of the algorithm before it is passed to dependency parser. Then the triplet extraction component gets the refined input and produces the triplets which contains the basic relationship of sentences. Then those triplets are saved in the local storage. Then the Q & A(Question and Answering) Module, searches all the triplets stored when a particular query is made, and retrieves all the possible answers to that particular question.

4. Implementation

4.1 Introduction

According to the high-level design and methodology given in the previous chapter, implementation modules were planned and carried out. Main objective of this work was to find a solution to the research problem, i.e “Finding a well-established mechanism to extract information from a given domain using a semantic approach”. By reviewing and referring previous research work, research gap has been identified and high-level design and methodology is formed after analysing the domain and deciding on a suitable approach for this work. This project is based on the domain of News Items of Music Artists collected from social media. As the main data collection medium, RSS Feeds/Twitter API are used. To create the sample test cases and for the evaluation, news related with the music artist are collected. By providing a RSS Feed URL as the input, system will collect the data from the provided Feed URL and, process the data and generate triplets as the output content. If the user does not provide a RSS feed URL as the input, default URL will be taken as the input. After generating triplets, end user can query the existing triplets and retrieve all possible answers related to the particular question. This Question Answer Module (Q &A Module) provides answers to the domain knowledge gathered by the triplets and it is capable of only providing answers to simple questions. With those two constraints given, implemented system is capable of collecting data from RSS Feeds and extract meaningful information from it, represent these data in a structured format and query any kind of simple question and retrieve the answer. Also, user can only view created triplets without performing a query in the Question Answering Module. In the following sections, it gives a complete description of the implemented solution.

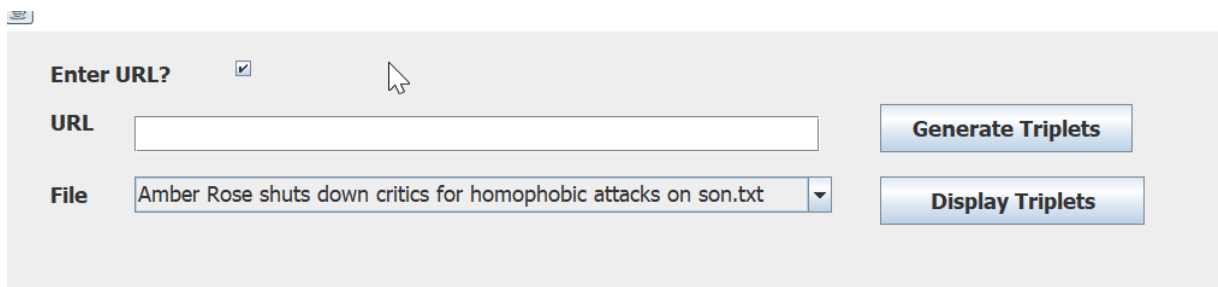
4.2 Input Process and Output

This sub section gives a detailed description of the system in terms of input, process and the output. This includes the user interfaces that has been designed and a brief description of the process.

Input

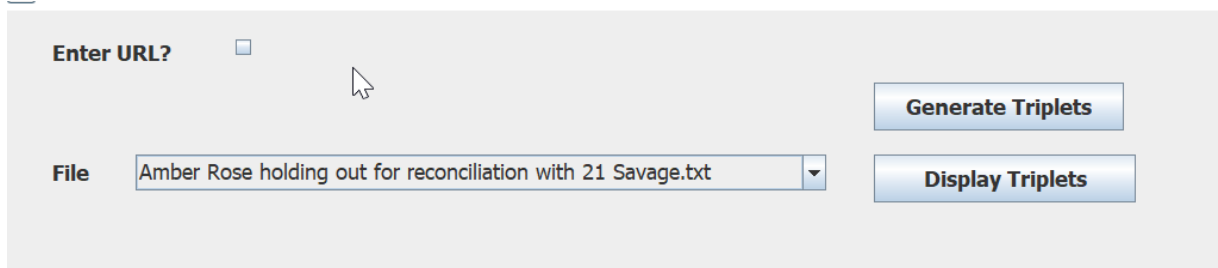
Input for the system can be a RSS Feed URL which has news related to music artists (celebrities) like <http://www.music-news.com/rss/UK/news?includeCover=true> or a link to a Twitter Music News public page. System allows the user to input a RSS Feed URL or else is allowed to continue with the default URL provided(RSS Feed URL-[25](http://www.music-</p></div><div data-bbox=)

news.com/rss/UK/news?includeCover=true). Below Figure 4-1 and Figure 4-2 shows how the design has provided the Enter URL option to the user.



The screenshot shows a web form with a checkbox labeled "Enter URL?" which is checked. Below it is a text input field for "URL" and a dropdown menu for "File" containing the text "Amber Rose shuts down critics for homophobic attacks on son.txt". To the right of these fields are two buttons: "Generate Triplets" and "Display Triplets".

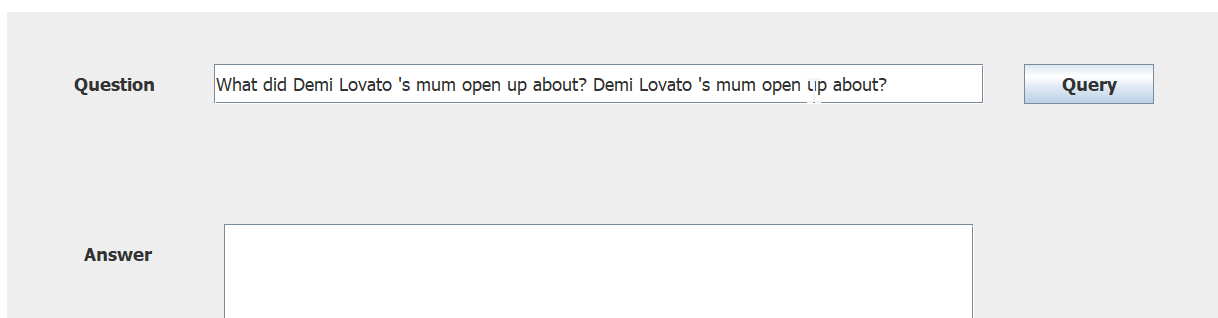
Figure 4-1



The screenshot shows the same web form as Figure 4-1, but the "Enter URL?" checkbox is now unchecked. The "File" dropdown menu now contains the text "Amber Rose holding out for reconciliation with 21 Savage.txt". The "Generate Triplets" and "Display Triplets" buttons remain on the right.

Figure 4-2

Apart from the basic input, after generating the triplets using the existing triplets available, user can select a particular data file which he needs to display triplets. By clicking on the Display Triplets button user can view the generated triplets. As shown is in **Error! Reference source not found.**, by selecting a Text file using the dropdown option these Triplets can be viewed. After generating triplets can go to the Q&A module by clicking on Question Generation button. Input for this module, is a question given in the text format as shown in below Figure 4-3.



The screenshot shows a Q&A form with a text input field for "Question" containing the text "What did Demi Lovato 's mum open up about? Demi Lovato 's mum open up about?". To the right of this field is a button labeled "Query". Below the question field is a larger text input field for "Answer".

Figure 4-3

Process

Following are the four main processes of this system.

- Retrieving Data
- Filtering and Pre-Processing Module

- Generating and Displaying Triplets
- Query Data (Question Answering Module)

Retrieving Data

When the input RSS Feed URL is given to this process, it retrieves all the RSS Feeds available, and for each RSS Feed it picks the title and the URL of the web page where the news contains. Then it fetches the URL and retrieve the content of the news article.

Filtering and Pre-Processing Module

After retrieving each news item, it filters and clean the unnecessary html content. After cleaning the content before processing the data by the triplet extractor, coreference resolution is performed to the set of data. When the coreference resolution is performed, most of the coreference are mapped into their real entities. To improve the accuracy of the extraction process coreference resolution should be performed as a part of the pre-processing task before passing the data in to the triplet extraction module. After this, an important pre-processing task performed is stemming and lemmatization. This is typically performed before or parallely with the triplet identification process.

Generating and Displaying Triplets

This process retrieves the stored data and pass the content to the Stanford parser to generate triplets. Then those triplets will be again saved as text files. Through display functionality, contents of the source text file and the output triplets is shown to the user.

Query Data (Question Answering Module)

When a user inputs the question on to the Question Answering module, it will query all the available triplets and retrieve all the possible answers.

Output

Main output of the system can be identified as the generated triplets. Outputs of each process will be listed below. Output of the Retrieving Data module will be a set of cleaned unstructured textual content. Filtering and Pre-Processing Module gives the resolved entities processed by the Coreference Resolution sub component. Generating and Displaying Triplets Module, generate the triplets and display the contents of the source text file and the output triplets. and

Figure 4-4 and Figure 4-5 shows the output of the three processes, Retrieving Data, Filtering and Pre-processing, Generating and Displaying Triplets.

Output of the Q & A module is the answer for the particular user's question that is generated related to the domain of the Music Artist News. Figure 4-6 shows the output of the queried question.

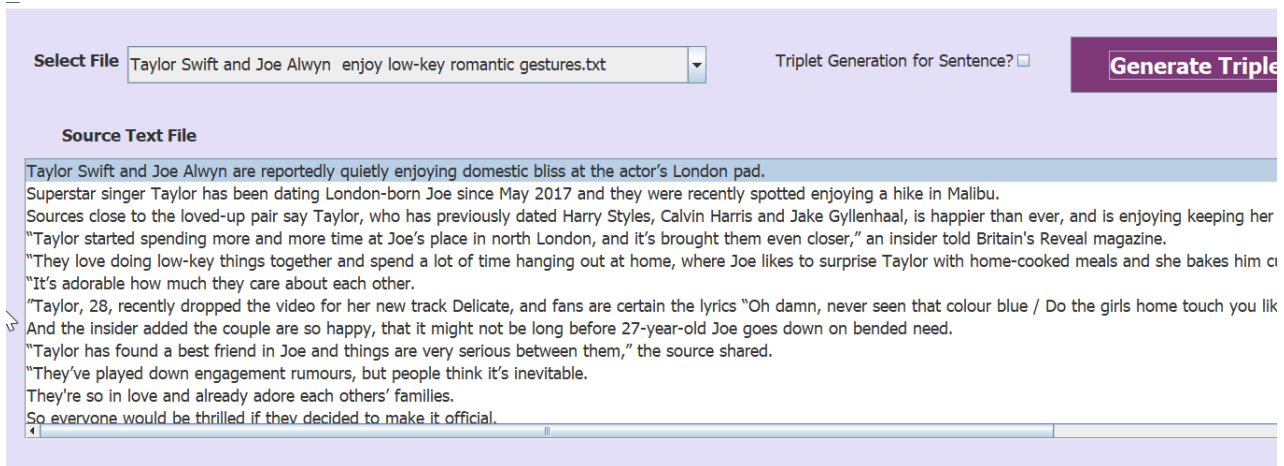


Figure 4-4

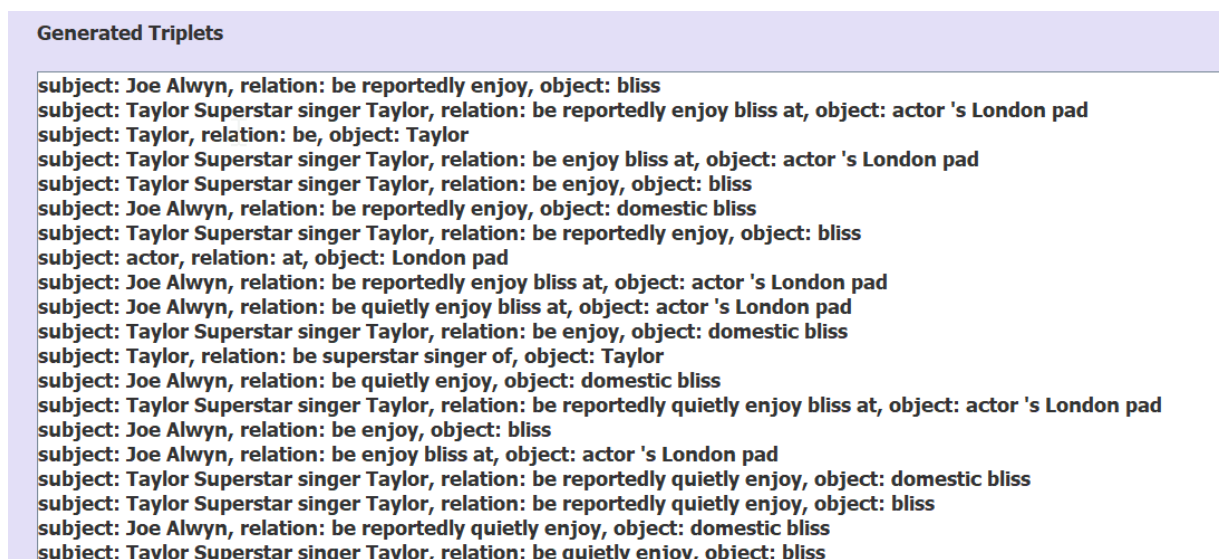


Figure 4-5

Question: Where did fifth harmony performed at?

Answer: Hard Rock Event Center in Hollywood on 11 May in Reykjavik

Figure 4-6

4.3 Basic Components of the System

This section describes the implementation of the system in detail. Implemented system consist of four basic components which can be introduced namely as, Data Retrieval Module, Filtering and Pre-Processing Module, Triplet Extraction Module and the Question Answering Module.

4.3.1 Data Retrieval Module

This Module is responsible for retrieving data from social media resources, mainly considering the RSS Feeds. As described in the methodology section, if the user does not specify the URL where to extract music related news, this component will get the default source as the system specified RSS Feeds URL. Figure 4-7 shows the interface design used to enter the URL and to retrieve data from RSS Feeds. Generate Triplets function involves with the first three modules of the system. For the ease of explanation and developing these modules, they were divided in to three parts as Data Retrieval, Filtering Pre-Processing and Triplet Extraction modules. This section only describes the Data Retrieval module.

Enter URL?

URL:

File:

Buttons: Generate Triplets, Display Triplets

Text File: [] Triplets: []

Figure 4-7

```

String urlstr46 = "http://www.music-news.com/rss/UK/news?includeCover=true";
ArrayList<String> as = null;
if (urlstr != null) {
    try {
        URL feedUrl = new URL(urlstr);
        SyndFeedInput input = new SyndFeedInput();
        SyndFeed feed = input.build(new XmlReader(feedUrl));
        // System.out.println("Feed Title: " + feed.getTitle());
        File f;
        FileWriter fw;
        int i = 1;
        String accessUrl = "";
        //Get each Feed from RSS Feeds
        for (SyndEntry entry : feed.getEntries()) {

            System.out.println("Entry Title: " + entry.getTitle());
            accessUrl = entry.getUri();

            f = new File("Text/" + entry.getTitle() + ".txt");
            fileName = f.getPath();
            fw = new FileWriter(fileName, true);

            //Write Data retrieved from RSS/Twitter
            GetWebData gb = new GetWebData();
            gb.getData(accessUrl, fw);

            fw.close();
            System.out.println("");
        }
    }
}

```

Figure 4-8

Following section provides a walkthrough of the implementation of the data retrieval module. Figure 4-8 shows the main part of the implemented module. When the user provides URL it passes the URL to the SyndFeed instance, which is capable of handling any type of Syndication input. Default URL is taken if the user does not specifies the input URL, and connect with the RSS Feed API (Rome API [24]). Then it instructs the SyndFeedInput to read the syndication feed from the input URL pointing to the feed [24]. After directing to the particular URL, each Syndication Entry is collected and the related news source relevant to each title is also fetched. For an example, for the RSS feed which has the title “Miley Cyrus and Ariana Grande rally with powerful March for Our Lives performances” and has the URL, [“http://www.music-news.com/news/UK/111964/Miley-Cyrus-and-Ariana-Grande-rally-with-powerful-March-for-Our-Lives-performances.”](http://www.music-news.com/news/UK/111964/Miley-Cyrus-and-Ariana-Grande-rally-with-powerful-March-for-Our-Lives-performances) Figure 4-9 shows the particular RSS Feed in the browser which is added as an extension.

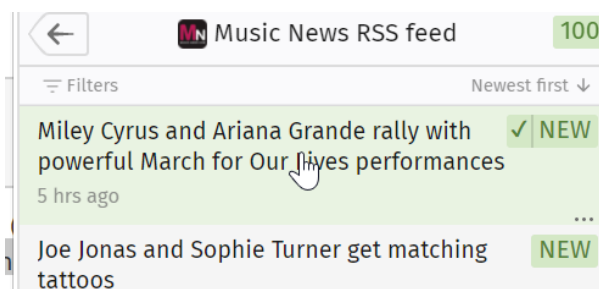


Figure 4-9

Then it accesses the URL of the web page of the particular news article. Then the URL and the Feed Title is passed to the Get Web Data sub module. Once the URL is passed to the connect method, it returns a Document object which fetches and parses the HTML document from the related webpage. Figure 4-10 shows how the data retrieved from the URL. Then the paragraphs of the HTML content are retrieved to the paragraph object and iterate through all the paragraphs of the news article and the text content is retrieved. Then the content in each article is written to a text file using the title as the previously fetched title from the RSS Feed. Figure 4-11 shows the sample text document where the retrieved data is passed.

```

public void webScrape(String stitle, String sUrl) {
    Document doc;
    try {
        title = "";
        content = "";
        title = stitle;
        String url = sUrl;
        doc = Jsoup.connect(url).get();

        Elements paragraphs = doc.select("p");
        Element firstParagraph;

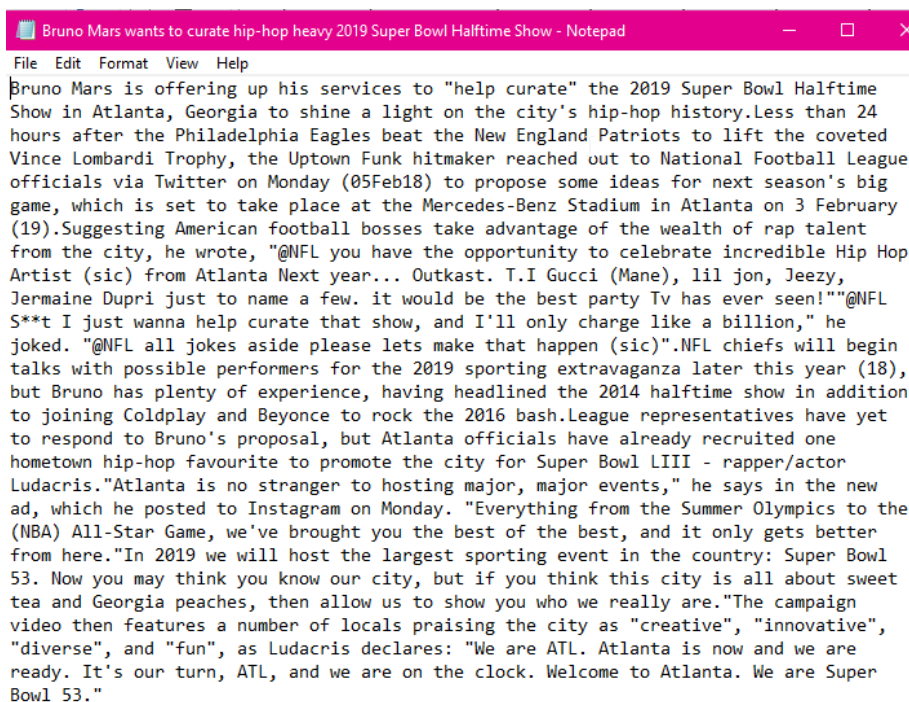
        for (int i = 0; i < paragraphs.size(); i++) {

            firstParagraph = paragraphs.get(i);
            String s = firstParagraph.text();
            Document dc = Jsoup.parse(s);
            System.out.println("DCCCC " + dc.wholeText());
            content += (dc.wholeText());

        }
        System.out.println(content);
        if (!content.isEmpty()) {
            writeData();
        }
    }
}

```

Figure 4-10



Bruno Mars wants to curate hip-hop heavy 2019 Super Bowl Halftime Show - Notepad

File Edit Format View Help

Bruno Mars is offering up his services to "help curate" the 2019 Super Bowl Halftime Show in Atlanta, Georgia to shine a light on the city's hip-hop history. Less than 24 hours after the Philadelphia Eagles beat the New England Patriots to lift the coveted Vince Lombardi Trophy, the Uptown Funk hitmaker reached out to National Football League officials via Twitter on Monday (05Feb18) to propose some ideas for next season's big game, which is set to take place at the Mercedes-Benz Stadium in Atlanta on 3 February (19). Suggesting American football bosses take advantage of the wealth of rap talent from the city, he wrote, "@NFL you have the opportunity to celebrate incredible Hip Hop Artist (sic) from Atlanta Next year... Outkast. T.I Gucci (Mane), lil jon, Jeezy, Jermaine Dupri just to name a few. it would be the best party Tv has ever seen!" "@NFL S**t I just wanna help curate that show, and I'll only charge like a billion," he joked. "@NFL all jokes aside please lets make that happen (sic)". NFL chiefs will begin talks with possible performers for the 2019 sporting extravaganza later this year (18), but Bruno has plenty of experience, having headlined the 2014 halftime show in addition to joining Coldplay and Beyonce to rock the 2016 bash. League representatives have yet to respond to Bruno's proposal, but Atlanta officials have already recruited one hometown hip-hop favourite to promote the city for Super Bowl LIII - rapper/actor Ludacris. "Atlanta is no stranger to hosting major, major events," he says in the new ad, which he posted to Instagram on Monday. "Everything from the Summer Olympics to the (NBA) All-Star Game, we've brought you the best of the best, and it only gets better from here." "In 2019 we will host the largest sporting event in the country: Super Bowl 53. Now you may think you know our city, but if you think this city is all about sweet tea and Georgia peaches, then allow us to show you who we really are." The campaign video then features a number of locals praising the city as "creative", "innovative", "diverse", and "fun", as Ludacris declares: "We are ATL. Atlanta is now and we are ready. It's our turn, ATL, and we are on the clock. Welcome to Atlanta. We are Super Bowl 53."

Figure 4-11

4.3.2 Filtering and Pre-Processing Module

After retrieving the content in the RSS Feeds, eliminating other available unnecessary noisy content and coreference resolution is performed. Most of the noisy data will be implicitly removed by the jsoup library and some of the noisy parts will be removed through this module. This component will mainly focus on the Coreference Resolution portion. Without performing the coreference resolution, lots of knowledge will be lost from the original content. When the coreference resolution is performed, most of the coreference are mapped into their real entities. To improve the accuracy of the extraction process, coreference resolution is performed as a part of the pre-processing task before passing the data in to the triplet extraction module. CorefResolution module of the implementation performs the Coreference Resolution task. Figure 4-12 shows a sample code fragment of the corefResolution module. Coreference resolution is invoked just before the triplet extraction, it uses the same pipeline object as of the TripletExtraction module. Figure 4-13 shows how the properties is set to the pipeline object and a detailed description of this will be provided in the triplet extraction module. How the coreference resolution is performed is previously described in the methodology section. In this module all the identifiable pronouns are mapped in to its original entity. And those pronouns will be replaced by the actual entity and the output sentence is passed as the input to the triplet extractor module.

```
private static String corefRes(String text) {  
    //StanfordCoreNLP pipeline = new StanfordCoreNLP(props);  
    Annotation doc = new Annotation(text);  
  
    pipeline.annotate(doc);  
    Entities which has coreferences, Camilla, She her etc. in a chain format  
    Map<Integer, edu.stanford.nlp.coref.data.CorefChain> corefs = doc.get(edu.stanford.nlp.coref.CorefCoreAnnotations.CorefChainAn  
    //Just the list of sentences ([sen1],[sen2]) in the type of CoreMap  
    List<CoreMap> sentences = doc.get(CoreAnnotations.SentencesAnnotation.class);  
  
    List<String> resolved = new ArrayList<String>();  
  
    for (CoreMap sentence : sentences) {  
        //tokens in the sentence with a number [Camila-1, Cabeio-2, is-3 etc..]  
        List<CoreLabel> tokens = sentence.get(CoreAnnotations.TokensAnnotation.class);  
  
        for (CoreLabel token : tokens) {  
            Integer corefClustId = token.get(edu.stanford.nlp.coref.CorefCoreAnnotations.CorefClusterIdAnnotation.class);  
            System.out.println(token.word() + " --> corefClusterID = " + corefClustId);  
  
            edu.stanford.nlp.coref.data.CorefChain chain = corefs.get(corefClustId);  
            //System.out.println("matched chain = " + chain);  
  
            if (chain == null) {
```

Figure 4-12

```
props.setProperty("annotators", "tokenize,ssplit,pos,lemma,ner,depparse,mention,coref,natlog,openie");  
pipeline = new StanfordCoreNLP(props);
```

Figure 4-13

4.3.3 Triplet Extraction Module

This is the main component of the implementation as it is responsible for extracting meaningful content from the collected data. After connecting with the relevant social media and retrieving, filtering and pre-processing data, the document which contains natural language text, will be directed to the Triplet Extraction module. This is directly connected with the main objective of this research work as this module extracts important information using the semantic analysis approach. To extract meaningful data, relationship between each sentence is identified and this is done through presenting the identified relationship using the triplet format. Triplets are extracted in the form of <Subject, Predicate, Object>. To extract those triplets, Stanford Dependency Parser and the Stanford Open IE Tools are used. As described in the methodology section to identify those triplets Stanford Open IE RelationshipTripletAnnotation algorithm is used. Figure 4-14 shows the code fragment used for the triplet extraction.

```
public static ArrayList<String> getTriplets(String sn) throws Exception {  
    // Create the Stanford CoreNLP pipeline  
    // creates a StanfordCoreNLP object, with POS tagging, lemmatization, NER, parsing, and coreference resolution  
    props.setProperty("annotators", "tokenize,ssplit,pos,lemma,ner,depparse,mention,coref,natlog,openie");  
    pipeline = new StanfordCoreNLP(props);  
  
    //System.out.println("");  
    sent = corefRes(sn);  
  
    //System.out.println(sent);  
    doc = new Annotation(sent);  
    pipeline.annotate(doc);  
  
    // Loop over sentences in the document  
    ArrayList<String> arrList = new ArrayList<String>();  
    for (CoreMap sentence : doc.get(CoreAnnotations.SentencesAnnotation.class)) {  
        // Get the OpenIE triples for the sentence  
        Collection<RelationTriple> triples  
            = sentence.get(NaturalLogicAnnotations.RelationTriplesAnnotation.class);  
        // Print the triples  
        for (RelationTriple triple : triples) {  
            arrList.add("subject: " + triple.subjectLemmaGloss() + ", relation: " + triple.relationLemmaGloss() + ", object: "  
//            System.out.println(triple.confidence + "\t"  
//            "subject: " + triple.subjectLemmaGloss() + ", relation: " + triple.relationLemmaGloss() + ", object: "
```

Figure 4-14

First it creates a StanfordCoreNLP object, with POS tagging, lemmatization, NER, parsing, and coreference resolution annotators. As defined in the SetProperty method parameters, before dependency parsing tokenization, sentence splitting, part of speech tagging, named entity recognition tasks are set to be performed. Then the coreference resolution is performed and sentences with resolved entities are passed as the input to this module. Then a document object is created with the content of the input (pre-processed content of one single news article) and pipeline is annotated with the above created document. Afterwards, it loops through each sentence of the created document and retrieves OpenIE triplets for each sentence. Thereafter,

by appending all the triplets to a single data structure, triplets of the complete article are retrieved. Then those triplets are saved on to the local storage.

4.3.4 Question Answering Module.

As described in the previous sections as well, this module is not directly involved with the main system. This QA module is mainly used for the demonstration and evaluation purpose of the system. Mostly this module is responsible for querying a particular user question and for the evaluation of the system. All the questions generated will be less complex and related with the “News Items related to Music Artists” domain. This evaluation module is developed to query two types of questions. They are namely Querying the Object and Querying the Subject (Questions based on subject or Object). Detailed description of the logic of this module is presented in the methodology section and the evaluation chapter. Figure 4-15 shows the important code sections of this module. When a question is received to this module it is converted in to triplet format.

```
public static ArrayList<String> queryTriplets(String ques) {  
  
    ArrayList<String> matchedValues = new ArrayList<String>();  
    System.out.println("Generated ques:"+ques);  
    ques = ques.split("\\?")[0];  
  
    writeAll();  
    String trp = readData(ques);  
    /      String trp = "subject: x, relation: do, object: soap opera";  
  
    //Changing the question format to string  
    String sentence = "";  
    sentence = ques.split(quesType)[1];  
    if (objType == "Thing" || objType == "Location" || objType=="Person Obj") {  
        sentence = sentence + " trp";  
    } else if (objType == "Person" ) {  
        sentence = "trp " + sentence;  
    }  
  
    sentence = sentence + ".";  
  
    System.out.println(sentence);  
  
    ArrayList<String> as = OpenIE.getTriplets(sentence);  
  
    System.out.println("Triplets..");  
    for (String a : as) {  
  
        System.out.println(a);  
        Triplet = a;  
    }  
  
}
```

Figure 4-15

First it is converted from question format to simple sentence format. When converting to the simple sentence format, question portion (Auxiliary Verb) is removed from the sentence. When the Auxiliary Verb is removed there will be a missing part of the sentence. This missing part of the sentence is identified and replaced with a constant (“trp”-which indicates a word). To replace the question portion, first the question type is identified and accordingly, the constant will be appended to the particular location. Figure 4-16 (a) and (b) shows conversion from question sentence into simple sentence format.

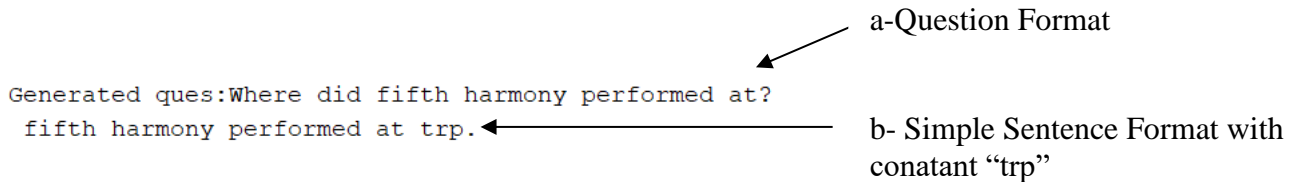


Figure 4-16

Then this complete sentence is sent to the triplet extractor module, which generates the following output in Figure 4-17.

```
fifth harmony performed at trp.
Triplets..
subject: fifth harmony, relation: perform at, object: trp
subject: harmony, relation: perform at, object: trp
```

Figure 4-17

By searching through all the extracted triplets matching against the generated triplet from the question, all the matches will be identified, and the non-matching portion will be retrieved as the output (Answer to the question).

4.4 Tools and Technologies Used

Java is used as the programming language to develop this system with the NetBeans IDE 8.2 as the Integrated Development Environment. Stanford Dependency parser and the Open IE is used to extract basic triplets and the entities of the given input sentence. Rome API is used to retrieve data from RSS Feeds and Jsoup library which is a java-based library which works with HTML based content. Following sub sections provides a basic overview of the tools and technologies used.

4.2.1 Rome API

ROME API is a Java based API, which provides a framework for working with RSS and Atom feeds. It's open source and licensed under the Apache 2.0 license. ROME includes a set of parsers and generators for the various flavours of syndication feeds, as well as converters to

convert from one format to another. The parsers can return Java objects that are either specific for the format you want to work with, or a generic normalized SyndFeed class that lets you work on with the data without bothering about the incoming or outgoing feed type [25].

Parsing Process

Rome is based around an abstract model of a Newsfeed or "Syndication Feed." Rome can parse any format of Newsfeed, including RSS variants and Atom, into this model. Rome can convert from model representation to any of the same Newsfeed output formats. Internally, Rome defines intermediate object models for specific Newsfeed formats, or "Wire Feed" formats, including both Atom and all RSS variants. For each format, there is a separate JDOM based parser class that parses XML into an intermediate model. Rome provides "converters" to convert between the intermediate Wire Feed models and the idealized Syndication Feed model.

Following is the basic process of Rome Newsfeed parsing,

- 1) When the SyndFeedInput object is created, it calls to the object to parse a Newsfeed.

```
URL feedUrl = new URL("file:blogging-roller.rss");
SyndFeedInput input = new SyndFeedInput();
SyndFeed feed = input.build(new InputStreamReader(feedUrl.openStream()));
```

- 2) SyndFeedInput delegates to WireFeedInput to do the actual parsing.
- 3) WireFeedInput uses a PluginManager of class FeedParsers to pick the right parser to use to parse the feed and then calls that parser to parse the Newsfeed.
- 4) The appropriate parser parses the Newsfeed parses the feed, using JDom, into a WireFeed.
- 5) SyndFeedInput uses the returned WireFeedInput to create a SyndFeedImpl. Which implements SyndFeed. SyndFeed is an interface, the root of an abstraction that represents a format independent Newsfeed.
- 6) SyndFeedImpl uses a Converter to convert between the format specific WireFeed representation and a format-independent SyndFeed.
- 7) SyndFeedInput returns to you a SyndFeed containing the parsed Newsfeed [25].

4.2.2 Overview of Jsoup Library

Jsoup is a Java based library to work with HTML based content. It provides a very convenient jsoup library provides following functionalities.

1. Multiple Read Support - It reads and parses HTML using URL, file, or string.
2. CSS Selectors It can find and extract data, using DOM traversal or CSS selectors.

3. DOM Manipulation It can manipulate the HTML elements, attributes, and text.
4. Prevent XSS attacks, can clean user-submitted content against a given safe white-list, to prevent XSS attacks.
5. Outputs tidy HTML.
6. Handles invalid data - jsoup can handle unclosed tags, implicit tags and can reliably create the document structure.

4.2.3 Overview of Stanford Core NLP (Which provides the dependency parser)

Stanford CoreNLP provides a set of human language technology tools. It can give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases and syntactic dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open-class relations between entity mentions, get the quotes people said, etc.

CoreNLP assist with common set of tasks as listed below:

- An integrated NLP toolkit with a broad range of grammatical analysis tools
- A fast, robust annotator for arbitrary texts, widely used in production
- A modern, regularly updated package, with the overall highest quality text analytics
- Support for a number of major (human) languages
- Available APIs for most major modern programming languages
- Ability to run as a simple web service [20].

4.5 Summary

This chapter provided a detailed description of the Implementation of the system which is implemented according to the high-level design and methodology presented. For the ease of the development, high level design of the system is broken down to several components accordingly and development is carried out in the following order. First the data collection and filtering are done and the output is passed to the Filtering and pre-processing module. By further filtering and cleaning, output is passed for the coreference resolution. This is performed by mapping all the pronouns to their real entities. Then output content of the above is passed to Triplet extraction module. This module identifies the binary relationship in the format of triplets. Above described “Data Retrieval”, “Filtering and pre-processing” and “Triplet Extraction” are the three basic components of the main system. Apart from the main system, Question Answering module is implemented for the Evaluation and demonstration purposes of

the main system. Brief description of the implementation of the system is covered in this section. Complete functionality of this is provided in the Methodology and the Evaluation chapters as this module belongs to the Evaluation section.

Through this implementation, main objective of this research work is accomplished, and a solid answer is provided to the research problem.

5. Evaluation and Testing

5.1 Introduction

This opening chapter provides a detailed description of the evaluation process which describes how evaluation and testing is performed to the implemented information extraction system including the test results. Data obtained for the evaluation is extracted mainly RSS Feeds. These data set was cleaned and prepared in a way that can be directly processed by the system. Data set for evaluation was chosen from the domain of “Music Celebrity News Articles”. This dataset was selected basically based on following reasons: fact-based nature of the news feeds, ease of preparation and testing due to short and article format of the feeds. Though this set of public data is not directly involved with the evaluation, it is used to identify and specify the knowledge of the domain which is given as the domain scope of the evaluation.

A separate Q&A (Question and Answer) module is developed for the evaluation process of the system, which allows the user to query against the extracted knowledge from the implemented system. Extracted data from the public social media, is used to identify the knowledge which is used to build a set of questions. Those questions are passed to the Question Answering module to retrieve the output results. User can check whether the knowledge is correct or incorrect by comparing the system generated answer and the user generated answer for a question.

5.2 Evaluation Model

As summarized in the introduction section, Evaluation of the implemented Information extraction system is conducted based on the developed Question Answer Module. With the aid of a set of users, human Testing is performed for the evaluation of the system rather comparing the extracted knowledge with a domain knowledge base or any other automated system.

Evaluation model of the system contains basic four steps as described below.

1. User Selection and Question Preparation:

First a set of eligible users were selected as the human testers of the system. These users were selected through accepting a given set of criteria: User has a sound English literacy where the particular user should be able to understand a given sample corpus in the domain of music celebrity information and generate quarriable set of questions and produce answers for the given set of questions. When preparing the set of questions another important point considered is, there are three types of questions generated to evaluate the system. As this Information Extraction System will generate triplets, answer for a particular

question would be, subject, object. Relation related queries could be generated by the combination of subject and object can be generated, but only the most frequent types of answer type for a triplet-based knowledge querying would be the first two types. So basic types of generated questions would be,

- 1) Querying the Object.
- 2) Querying the Subject.
- 3) Querying the relation (very rare)

2. System Evaluation,

It queries the given question against the QA (Question Answer) module and generate the output. This evaluation module is developed to query above two types of questions (Questions based on subject or Object). Before generating the output for a user's generated question, the system should contain the knowledge of the extracted content. In order to do that an algorithm is developed to retrieve all the existing triplites, extracted from the RSS Feeds. Then the querying algorithm is developed to query the user generated question. In the querying algorithm, first the question is pre-processed and converted in to a triplet format then the generated triplet is queried against all the existing extracted triplets. By comparing with the existing triplets, the result related with the user query is retrieved as the output.

3. User Evaluation

This consist of user reviewing the particular set of given corpuses and answering the questions. This step can be performed in the first step of the evaluation as well. When answering the questions, questions were shuffled between users so that questions wouldn't be distributed to the same set of users who generated them.

4. Comparison of the two Evaluation Results

Comparison of the user generated answers and the system generated answers. In this step if there are complex words or words which contain out of the data set is eliminated first. Then the answer is compared with the system generated answer. A detailed description of this step and the results of the comparison is given in the Results section of this chapter.

5.3 Test Data for the Evaluation

Data scraped, extract from web sources pointed by the RSS Feeds are collected, cleaned and pre-processed to build the Extraction system and the extracted data are evaluated with the assist of above described section. This section outlines the test data and the sample test scenarios which is used to evaluate the implemented system.

After collecting data from web sources directed by RSS Feeds data is cleaned by removing unwanted html tags and other unnecessary textual data. Then the set of user generated questions are being prepared as the test data. For the evaluation purpose of this system was tested against the extracted knowledge from 100 corpuses. User was allowed to choose 10-20 different text files and generate 15-20 questions from the knowledge in the input files. Then the questions were distributed within 10-15 users with the set of sample source files to answers for the question using the domain knowledge of the input data files.

Figure 5-1 shows a sample set of such questions generated by the user, and Figure 5-2 shows the set of input data sources which is used to generate question.

What did Demi Lovato had?
Who seek rehab help?
Where did fifth harmony performed at?
Who has a song writing skill?
Who recently drop a video?
What Swift is accuse for?
Who cut out alcohol?
What did Rihanna post?
Who does soap opera?
From whom is Cheryl inspired by?
Where does Hollywood Bowl located at?
With whom veteran Welsh crooner is a close pal?
Who is happy with her love life?
Who canceled the North American tour?
Why Cardi B turned off Twitter?
Who celebrated as recieved green card?
Who likes Taylor Swifts Music?

Figure 5-1

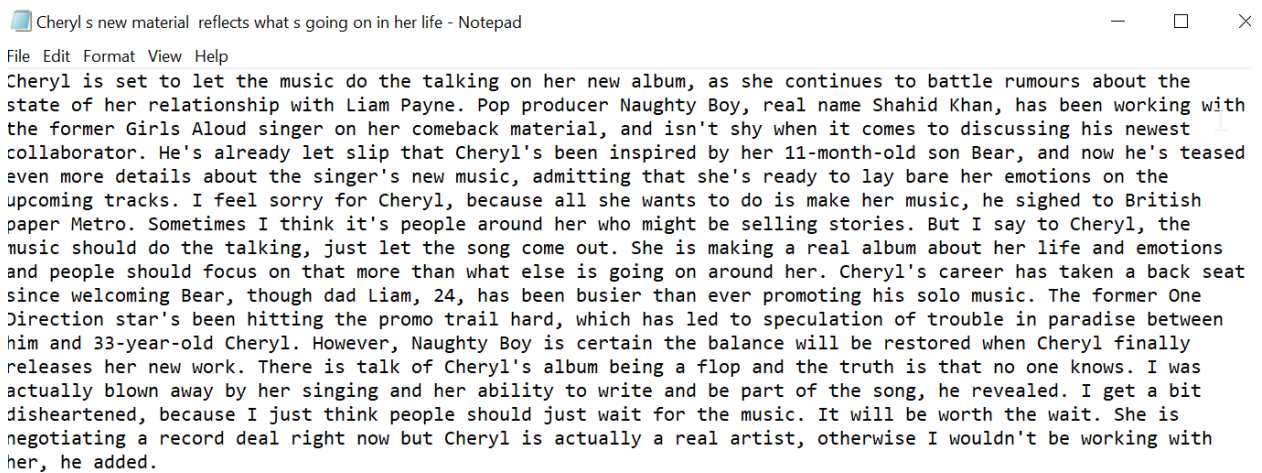


Figure 5-2

Above set of questions are generated using the guidelines given in the evaluation model component. Following are two sample types of questions generated as the given question formats.

Sample Sentence:

“Katy Perry praised Taylor Swift's songwriting skills as she responded to an American Idol contestant's apology for being a fan of the Bad Blood star.”

User Generated Question:

Who praised Taylor Swift's songwriting skill?

5.4 Results and Discussion

This section describes test results generated by the system, the accuracy of each test scenario and discuss on the overall evaluation of the system. By analysing the tests results it estimates an accuracy percentage for the Implemented Information Extraction System.

As described in the previous section there were 10 test scenarios, where a set of questions are given to a human user and generate and compare answers of each scenario and evaluate the scenario assigning the system output a percentage. Figure 5-3 **Error! Reference source not found.** shows the percentage of the correct answers generated by the system. Overall accuracy calculated for the system is 57.34% . Though the system is unable to generate accurate triplets, it is capable of generating triplets which contains most of the important information from the input data.

When considering the accuracy of the each of the system, triplet extraction module has a accuracy of 71.2% and the Coreference resolution shows the accuracy of 43.4%. Though

coreference Resolution has a low rate of accuracy, it enhance the extraction capacity of the system by considering a large part of entities which is typically being discarded. Without the Coreference resolution module, it will limit the capacity of the extraction system as it discards some entities which contain valuable information.

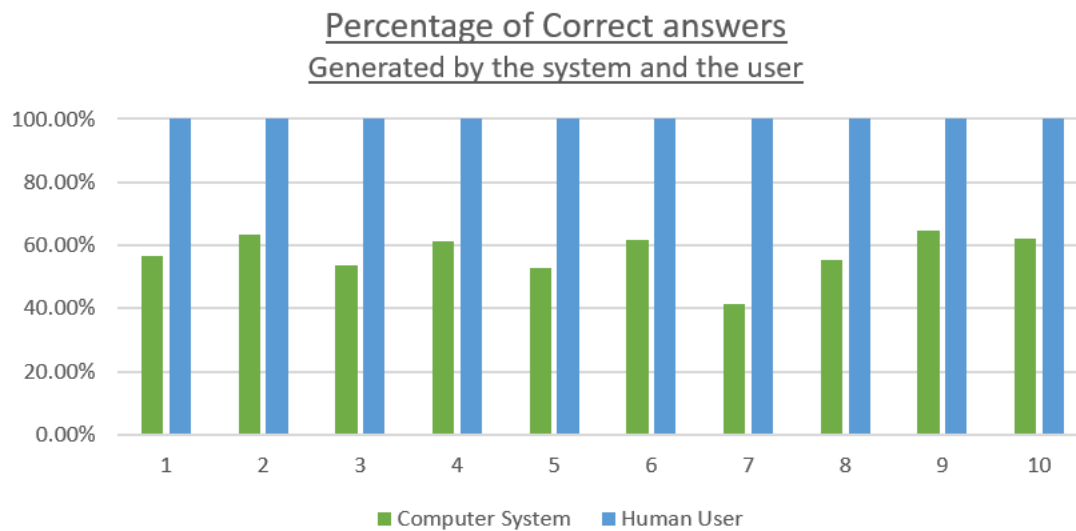


Figure 5-3

5.5 Summary

This chapter outlines the evaluation of the implemented information extraction system, how testing is conducted and discussed on the test results and overall accuracy of the system. Evaluation model of the system is divided in to four basic levels which is User Selection and Question Preparation, System Evaluation, User Evaluation and the Comparison of the two evaluations. By performing this evaluation to data extracted from 100 corpuses (Nearly 6000 triplets) by generating 200 questions and comparing system generated answers and user generated answers, these results shows 67% of accuracy for the system. By this evaluation it clearly shows that the research objective: To build an efficient and powerful information extraction algorithm, which analyse the information semantically in a digital text and map the content into a machine readable, meaningful, structured format.

6. Conclusion and Future Work

6.1 Introduction

Previous chapter discussed how the evaluation is performed for the implemented information extraction system, on the test results and discussed on the accuracy achieved by the system. This chapter concludes this thesis by discussing the summary of the work that has been completed. It further discusses on research findings, problems encountered in the designing and implementation phases of the system, limitations and finally the advancements that can be added to the presented work.

6.2 Research Findings

Implemented system was inspired through the many previous work done in the research field of Information Extraction. But it was mainly motivated through the Knowledge Extraction Module of the “An Ontology Based Natural Language Story Generation Approach” research work [1]. By performing a thorough literature survey, many approaches were analysed, and research gap was identified. Many of the implemented approaches has covered only basic tasks like syntactic information (POS tagging, chunking, parsing etc.), semantic role labelling or named entity recognition etc. Approaches introduced in area of semantic information extraction is limited Most of them are domain specific approaches and lack of extracting important meaningful content. Further, there are many high complex approaches that has been introduced, are domain specific and time consuming and requires lot of training. By analysing the research gap, main research problem was identified. “There is no well-established mechanism to Extract Information Semantically from Raw Textual Data and Discovering Valuable and Relevant Information from it”. To find out a solution for the research problem, common gaps in previous approaches were identified and the high-level design was implemented. This design was implemented based on introducing a general and simple approach which can identify meaningful information from any kind of domain. But for the ease of implementation, testing and evaluation, the work has been limited to the domain of Music Celebrity News Articles.

Through this research work, a novel approach has been introduced for extracting meaningful content by identifying relations from plain text. These relations are arranged in the format of Triplets. Triplets are in the form of subject, predicate and object. So, given a sentence, core meaning of that sentence is extracted, hence the most important knowledge of the sentence is captured through these Triplets. According to the evaluation section, results of the evaluation shows 67% of accuracy for the system by testing data extracted from 100 to 150 corpuses. As

such, through this implementation, main objectives of this research work are accomplished. Thus, a solid solution is provided to the research problem.

6.3 Problems Occurred

When designing and implementing the system, there were several significant problems that were occurred. This section provides a detailed description of those problems and, how they were handled as far as possible. Initially the data was gathered on domain of the fan-based opinions of Music Artist from social media. When those data have been gathered up to a certain extent, a significant problem that occurred was that the data collected were not factual. They were more people biased, hence information extraction from such data was pointless. So, data collection was slightly modified without any alteration on the scope- which was Music Artist related data and their Work (Tracks). Therefore, issue with the data collection was solved by collecting data from Music Artists Related News retrieved via RSS Feeds or Twitter.

When performing the Triplet Extraction and coreference resolution, there were duplicate triplets were produced due to the multiple relation extraction. This could not be solved as this occurs when dealing with natural language and there are occurrences where one single sentence would be represented in multiple ways.

After identifying the triplets, to extract information related with the Music domain was a challenge. Building up a rule-based approach to filter the domain is a tedious task, as the set of rules which contains domains contains lots of information. As there are limited well-established existing knowledge bases related with the Music domain, extraction component had to be limited only for relation extraction.

6.4 Limitations

Approach that has been introduced is capable of identifying the semantic relation of the extracted entities. As this system has presented a simpler approach, information extraction process is developed in a shallow level and does not provide any pragmatic kind of information. Also, as described in the problems section when performing the triplet extraction and coreference resolution, addressing issue of the duplicate triplet generation could not be resolved. This was due to multiple relation extraction. when dealing with natural language where there are occurrences that one single sentence would be represented in multiple ways. Eliminating this limitation would be an opening for another research problem thus it is not considered in this work. When extracting meaningful content using a deep analysis approach, recognizing the domain and based on the identified domain, performing a semantic information extraction task would give more successful than the introduced approach. Also, it will be more

expensive and mostly open multiple research ideas, which will require more hands-on experience on this field.

6.5 Future Work

As declared in the limitation section, this approach can be enhanced by using a domain knowledge base and, extracting required information with the assist of the domain knowledge base. So, there will be much more meaningful content rather extracting information in a shallow level. As an alternative to building up a rule-based domain an ontology could be used related to the music domain. By using the above approaches or any other relevant approaches, this work can be enhanced as an ontology-based information extractor. Further, by introducing hierarchical named entity recognition task, this approach can be enhanced along with the relationship extraction module. Multiple triplet generation issue can be eliminated, and accuracy of the output can be increased. Further, by merging the existing co reference resolution task with a rule-based approach, accuracy of the co reference resolution could be enhanced. Also populating a knowledge base using the extracted content, this system can be used as the input for a natural language generator system.

References

- [1] K. Kodikara, M. M. F. Shafna, M. R. F. Nazla, B. P. P. Chandrasena, J. L. Amararachchi, and D. N. Koggalahewa, “An Ontology Based Natural Language Story Generation Approach : WELOTA,” in *NCTM - SLIIT*, 2014, vol. NTCM2014;0, no. 18003591, p. 8.
- [2] C. C. Aggarwal and C. X. Zhai, *Mining text data*, vol. 9781461432. 2013.
- [3] “GATE.ac.uk - index.html.” .
- [4] “RiTa.” [Online]. Available: <http://rednoise.org/rita/index.html>. [Accessed: 22-Aug-2014].
- [5] H. Alani *et al.*, “Automatic Ontology-Based Knowledge Extraction from Web Documents,” *IEEE Intell. Syst.*, vol. 18, no. 1, pp. 14–21, 2003.
- [6] “WordNet Search - 3.1.” .
- [7] M. B. Habib and M. Van Keulen, “Information Extraction for Social Media,” *Work. Semant. Web Inf. Extr.*, pp. 9–16, 2014.
- [8] A. Ritter, Mausam, O. Etzioni, and S. Clark, “Open Domain Event Extraction from Twitter,” in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 1104–1112.
- [9] I. Mani and G. Wilson, “Robust Temporal Processing of News,” in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 2000, pp. 69–76.
- [10] P. Knees and M. Schedl, “Towards Semantic Music Information Extraction from the Web Using Rule Patterns and Supervised Learning,” *Music Recomm. Discov.*, pp. 18–25, 2011.
- [11] “Last.fm - Listen to free music and watch videos with the largest music catalogue online.” .
- [12] M. DeGagne, “AllMusic | Record Reviews, Streaming Songs, Genres & Bands,” *allmusic.com*, 2017. [Online]. Available: <https://www.allmusic.com/>.
- [13] E. Zangerle, M. Pichl, W. Gassler, and G. Specht, “#nowplaying Music Dataset,” in *Proceedings of the First International Workshop on Internet-Scale Multimedia Management - WISMM '14*, 2014, pp. 21–26.
- [14] “Availability - Spotify.” [Online]. Available: <https://www.spotify.com/>. [Accessed: 07-Jul-2018].
- [15] R. Anantharangachar, S. Ramani, and S. Rajagopalan, “Ontology Guided Information Extraction from Unstructured Text,” vol. 4, no. 1, pp. 19–36, 2013.
- [16] A. Maedche, E. Maedche, and S. Staab, “The TEXT-TO-ONTO Ontology Learning

- Environment,” in *Software Demonstration at ICCS-2000 - Eight International Conference on Conceptual Structures*, 2000.
- [17] G. Neumann, R. Backofen, J. Baur, M. Becker, and C. Braun, “An Information Extraction Core System for Real World German Text Processing,” *Proc. ANLP97*, p. 9, 1997.
- [18] P. Cimiano and J. Völker, “Text2Onto: A Framework for Ontology Learning and Data-driven Change Discovery,” in *Proceedings of the 10th International Conference on Natural Language Processing and Information Systems*, 2005, pp. 227–238.
- [19] Stanford NLP Group, “The Stanford Natural Language Processing Group,” 2016. [Online]. Available: <https://nlp.stanford.edu/projects/coref.shtml>. [Accessed: 25-Mar-2018].
- [20] Stanford NLP Group, “The Stanford Natural Language Processing Group,” 2016. [Online]. Available: <https://nlp.stanford.edu/software/nndep.shtml>. [Accessed: 24-Mar-2018].
- [21] J. H. M. Daniel Jurafsky, “Dependency Parsing,” in *Speech and Language Processing*, 3rd ed., 2017, p. 128.
- [22] M. Marcus, A. Taylor, R. MacIntyre, A. Bies, M. Cooper, Constance Ferguson, and A. Littman, “Penn Treebank P.O.S. Tags.” [Online]. Available: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html. [Accessed: 24-Mar-2018].
- [23] Stanford NLP Group, “The Stanford Natural Language Processing Group,” 2016. [Online]. Available: <https://nlp.stanford.edu/software/openie.html>. [Accessed: 24-Mar-2018].
- [24] “ROME - Home,” 2017. [Online]. Available: <https://rometools.github.io/rome/>. [Accessed: 08-Feb-2018].
- [25] “ROME - How Rome works.” [Online]. Available: <https://rometools.github.io/rome/HowRomeWorks/index.html>. [Accessed: 26-Mar-2018].

