Multi-tool Methodology for Converting Text into RDF Triples

by

Elham Mahamedi

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Software Engineering and Intelligent Systems

Department of Electrical and Computer Engineering
University of Alberta

Abstract

There is no doubt that Internet becomes one of the most important sources of information. At the same time the amount of information stored on the web and available for users becomes enormous. In order to make this information more accessible and create prospects for software to process it automatically, a different format of storing information has been proposed by World Wide Web Consortium – it is called Resource Description Framework (RDF). This format can be described as a triple: subject-property-object. Application of RDF leads to creation of a highly interconnected network of nodes containing pieces of information. RDF could change how information is stored and processed on the web.

However, one of the most common formats of representing information on the web is and will be a simple text. A textual format is the most natural way of representing information used by individuals. Therefore, in this thesis, we focus on the task of translating text into RDF. The proposed approach is based on a combination of well-known Natural Language Processing tools: parsers, with a web-based tool for disambiguation, and our own algorithms for combining the results obtained from these tools and converting them into RDF triples.

Dedication

To my family

For their love, dedication and encouragement

Throughout my life

Acknowledgement

I would like to express my special gratitude towards my supervisors Dr. Marek Reformat and Dr. Witold Pedrycz for their kind support, valuable guidance and friendly advice throughout my educational career.

Table of Contents

List of Tables

List of Figures

1. Introduction

The web is a vast repository of distributed data while it grows with an astonishing rate. Dependency of users on the web becomes more and more pronounced. The variety of information available and stored on the web becomes a potential problem of how to access and process data. It has become more and more evident that the current way of storing data on the web would not lead to solutions of the issues raised above. On top of that, a large amount of data that is currently available and will be stored on the web is in the textual format. Text is the most natural form of expressing ideas and communicating between people. Therefore, textual data and information will be present on the web for the foreseeable future.

The above-mentioned issues create challenges of capturing essence of textual documents and translating them into a machine-readable representation. This task is not trivial since the information process should be able to preserve semantics of the text, as well as its grammatical structure. We propose here – via combining Natural Language Processing tools such as parsers and disambiguation techniques with our own algorithms – an approach for converting plain text documents into a machine-readable format.

1.1 Related Research Areas

Information extraction (IE), a form of natural language analysis, is the activity of automatic translation of unstructured information – such as a text – into structured information. Sometimes, this process is called "text analytics" [1]. IE involves annotating the unstructured text with entities, relations between entities while also extracting semantic relations between entities in text [1]. The relation between two specific entities can be learned from contexts. The accuracy of entity relation discovery is highly dependent on how accurate the entity resolution is during the annotation process [2].

Entity Recognition or Named-entity recognition, which is a key part of IE, identifies elements in text, and classifies them into predefined categories including universally accepted ones (i.e., individuals, organizations, locations) and other various categories such as times and dates [3].

The goal of IE is to produce a machine-readable structure of information in order to build and extend knowledge base and ontologies [3]. In order to store pieces of information for automatic querying and processing, IE gathers information from natural language texts. To store and process such data, the Semantic Web (see below) provides particular formats and standards [4]. In other words, while IE aims to extract relevant information from natural language texts, the Semantic Web provides a data representation format – Resource Description Framework – which is the basic data model used for building the Semantic Web [4].

Semantic Web represents a new paradigm of storing and processing data of the web that is highly dependent on the semantics of the data [4]. Unlike machines that are able to process large amounts data, humans as the main users of the web are innately incapable of finding a specific set of data when faced with too much information [5]. The data format used so far – such as HTML – is not suitable for content-sensitive machine processing due to a number of reasons such as: the lack of ability to reason about the meaning of data; the presence of certain ambiguities; and the lack of background knowledge for data processing and analysis [5]. Therefore, the Semantic Web provides a way of representing information distributed across the Web in a manner that is interpretable by machines [5].

IE does not provide any standard formats for representing target structures or storing extracted information. However, the Semantic Web comes with a very well defined data representation format called Resource Description Framework (RDF). It can be used to store extracted facts and to denote the target structures [4].

RDF ("Resource Description Framework") is a standard for encoding metadata in the environment of the Semantic Web [6]. In a more general way, the term metadata refers to any information expressed in RDF or similar formats; while in a stricter sense it also refers to data about documents such as title, author, and language of a text [4]. RDF is also defined as an abstract model, which converts knowledge into separate pieces [6].

Resource Description Framework is defined through the concepts of resources and properties, which are founded on the notion of making statements about resources in the form of *subject-predicate-object* expressions known as triples. A resource refers to anything that can be shown by a Uniform Resource Identifier (URI). A resource could be a part of or complete Web page, or

any physical object. Any relationship between the subject and object is also expressed by the predicate that is associated with a URI that points to a place, on the web, that contains its definition. The object could be considered as a literal or an alternative resource represented by a URI.

For instance, if there is a text file includes “James Gosling developed JAVA”; the extracted information from this text in the form of an RDF triple is as follows:

Subject: James Gosling

Predicate: develop

Object: JAVA

A graph-based model of this information in the form of a RDF triple is depicted in Figure 1.

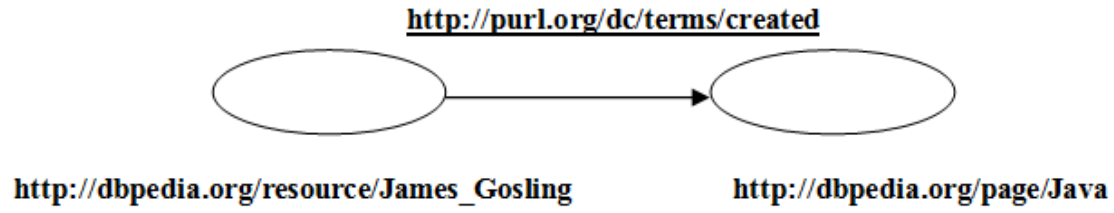

Figure 1. An example of an RDF triple: subject-predicate-object

While searching for “James Gosling” could render results both related and unrelated to the James Gosling who developed JAVA; RDF allows us to remove this ambiguity, by restricting the results by the URI <http://dbpedia.org/resource/James_Gosling >. Additionally, the subject <http://dbpedia.org/resource/James_Gosling > could be a subject of other RDF triples that are used to describe James Gosling, or an object of yet another RDF triples are descriptions of other resources.

1.2 Motivation

Documents are made up of statements related to specific subjects. However, these sources of valuable information do not have a sufficient structure for machines to be able to exploit them in the case of automation of inquiries. Consequently, the available growing pool of data would be

inaccessible and almost useless unless meaningful, related and useful information can be extracted from it [7].

To this end, the goal of this study is to automatically extract meaningful, simple and well-organized information from unstructured text. This approach is not domain specific. Hence, it is applicable to any activities such as providing results to search queries and creating knowledge bases. In this study, we used three different language processing tools including AlchemyAPI [8], C&C-Boxer [9], and Stanford-Senna (T2R) [10]. These tools, being a part of Natural Language Processing (NLP) methods, help us extract essential parts of information. This information, which consists of relationships and entities extracted from unstructured input texts, is used to create machine-readable information in the form of Resource Description Framework (RDF) triples. This transforms a difficult task of expressing useful and domain specific information into a simple task. Therefore, by utilizing these three NLP tools, we do not only rely on the grammar or syntactic structures of the processed sentences but also on their conceptual aspects or semantics. A text paragraph from any domain, and more precisely individual sentences of this paragraph, constitutes an input to these mentioned-above tools. In the case of complex sentences, we divide them into sentences with simple grammar structures by using special patterns explained in the prepossessing steps of this study. This division is done in order to ensure that all sentences lead to generation of RDF triples.

We used two different NLP tool (C&C-boxer and Stanford-Senna (T2R)) to determine the level of confidence in the generated triples. This is accomplished by showing that the RDF triples generated by different tools from the same sentences are similar and cover the essence of these sentences. We also use Stanford Corefrence-chain [11] tool in order to identify the references of any pronouns existing in the processed sentences. All this means that we process the paragraph thoroughly, and translate it to make well described and meaningful RDF triples while decreasing any ambiguity hidden in the text.

1.3 Outline of the Thesis

In order to achieve the above-mentioned goal, the remainder of this thesis is organized as follows. The second chapter provides a background and a summarization of some related work. In chapter 3, details of our contributions are described. This chapter provides the descriptions of

a pre-processing stage, processing of the results obtained from NLP tools, and a way of building RDF triples based on these results. Chapter 4 describes an evaluation process we proposed and applied in this study. It provides an overview of defined evaluation criteria. It contains the results of the evaluating process on a number of random sentences. The last chapter concludes the thesis by summarizing the work done in the thesis along with the main contributions. This chapter also includes some interesting areas for future consideration.

2. Background and Related Work

2.1 Text Analysis

The vast majority of valuable information existing online is in the form of unstructured textual data. These texts consist of information that can be extracted via text analysis processes including syntactic and semantic examination. Textual content analysis seeks to develop systems to reveal meaningful information. In these processes, ambiguity and fuzziness seen in natural languages requires a special treatment.

However, it is evident that regardless of its unstructured and vague characteristics, natural language is still the most convenient means of exchanging information. Despite the difficulty it faces and a lack of guarantee of successful elimination of vagueness from languages the text analysis remains an interesting and important task [12].

The most complicated and challenging part of analysis of natural languages or text mining is the presence of ambiguous grammatical rules. Different statistical techniques, artificial intelligence methodologies and modeling linguistic patterns are general methods overcoming this complication [12].

Textual documents have different syntax and semantic structures and their understanding requires various analyzing methods. These methods cover variety of syntactic as well as semantics aspects that express the actual meaning of the text. Therefore, multiple different Natural Language Processing tools are utilized to syntactically and semantically analyze textual data.

2.2 Natural Language Processing Tools

As explained in the previous section, to obtain semantic and syntactic information from any unstructured text the NLP techniques are utilized. Entities or named entities are interesting and useful information existing in any text. Most of the information contained in the text is related to these entities. Therefore, finding named entities in the text is one of the important steps to

uncover hidden information and create a structured version of this text in the form of RDF triples. In order to identify entities, a number of systems that use NLP techniques exist. AlchemyAPI [8] is one of such systems. It can recognize any named entity existing in the text. Another NLP based system tool that relies on syntactic and semantic forms is the C&C-Boxer tool [9]. C&C is a syntactical parser, while Boxer uses the output of C&C and performs a deep semantic analysis. Therefore, this tool is a suitable choice for processing sentences and extracting semantic and syntactic information from an unstructured textual data. T2R [10] is another semantic and syntactic analyzer system utilizing Stanford [13] and Senna [14] tools to analyze input text. This tool is able to translate information extracted from sentences in the form of RDF triples.

In this study, we design and develop a system to extract knowledge embedded in textual documents. To this end, we utilize the three mentioned-above tools which address major needs of the project, i.e., extracting syntactical information and semantics from sentences and converting it into a machine-readable format of RDF structures.

2.3 Related Work

The purpose of creating structured data from unstructured text can be achieved using various approaches. These approaches usually include annotation and processing of text using named entities and relations. This section describes a few selected examples. Each of the presented methods contains a different approach of creating RDF triples. There is also an interesting study which aims at a comparison of existing tools in the area of information/knowledge extraction.

In the LODifier [15] study, a number of tasks are combined in order to extract named entities and relations between them from a text. These entities and relations are eventually converted into an RDF representation that provides links to DBpedia and WordNet. The system performs a deep semantic analysis by assigning Wikipedia links recognized by wikifier to each named entity. Also, individual words are disambiguated via LODifier using UKB that performs a graph-based word sense disambiguation. The UKB outputs are converted to RDF WordNet URIs. This study creates RDF representation of the whole unstructured text, regardless of its topic. It uses C&C parser that is a statistical parser using combined categorical grammar. It also uses the parsing result as an input for Boxer that produces discourse representation structures (DRSC) (relevant

entities) and models the meaning of text in terms of relevant entities and relations between them. For the evaluation part, LODifier is applied to assess similarities between documents in newspapers and news existing in a TDT benchmark dataset. Similarity measures included measures without structural knowledge and those with structural knowledge. For the assessment, 183 positives pairs of documents with the same topics, and 183 negative ones with different topics were selected [15].

Asknet [16], similar to LODifier uses a processing method that is designed to automatically generate a semantic network. Asknet uses C&C and Boxer to determine semantic relations. Similarity scores are calculated using spreading activation to conclude which nodes refer to the same entities, and are mapped accordingly.

P. Exner and P.Nugues [17] introduces a system that takes unstructured Wikipedia texts as an input and generates RDF triples. The foundation of its text processing unit is semantic parsing of a text using Propbank (The Proposition Bank) [18]. This allows for annotating the text with predicates and arguments. The system uses co-reference solver which attempts to find the mentions in input text that refer to the same entity – this is accomplished by detecting noun phrases, and deciding which noun phrases are co-referential. In this study, both named entities and subsequent mentions to corresponding DBpedia URIs are linked. In order to form DBpedia RDF triples, an ontology-mapping module is used to map the predicate and different argument roles defined in PropBank Format onto a more general role sets using DBpedia properties. This module also matches the subject and object of triples to existing triples in DBpedia dataset. In situations when triples created with PropBank do not have equivalent triples in DBpedia, generalized triples are created. The DBpedia RDF triple is identified, and "PropBank triples" are linked with it. Additionally, new triples are added to the DBpedia dataset. For the evaluation section, 200 sentences were randomly selected to manually find subject-property-object structures. These structures were eventually compared to the corresponding triples extracted by this approach.

Ramakrishnan et al. [19] demonstrate a rule-based method for extracting the often-occurring complex entities in biomedical texts, define the relationships between them and finally convert those relationships into RDF triples. As the result, generated RDF triples are used to discover knowledge from text. It happens via locating paths comprised of the extracted relationships.

In the study by Gangemi [20], a few different Knowledge Extraction (KE) tools are compared based on possible functionalities they provide. The comparison is performed based on such aspects as "Named Entity Recognition" and "Terminology Extraction". The measures: precision "p", recall "r", and accuracy "a" were applied for the evaluation purposes. The input text for this study was originated from an online article of The New York Times. "Named Entity Recognition" was evaluated only for named entities represented as individuals in an ontology. The named entities or property names were also evaluated in the terminology extraction and resolution measures. In this research, the compared tools included: AIDA [21], Alchemy [8], Apache StanbolCiceroLite [1], DBpedia Spotlight[2], Fox[3], FRED[4], NERD[5], Open Calais[6], Wikimeta[7], and Zemanta[8]. They were evaluated from the point of view of recognizing named entities. Alchemy, AIDA and Zemanta had outstanding measures from this aspect. The precision "p" and accuracy "a" for both Alchemy and AIDA were 1.00 and 0.89 respectively. The same measures for Zemanta were 0.92 and 0.93. For terminology extraction parameter, which is based on class induction and property induction, five tools were compared including Alchemy, CiceroLite[9], FOX, FRED, and Wikimeta. Among them FRED had the highest scores. The precision "p" and accuracy "a" for FRED were 0.93 and 0.90 respectively. It should be emphasized that FRED is used to automatically generate RDF/OWL ontologies and Linked Data from a text.

[1] http://dev.iks-project.eu:8081/enhancer
[2] http://dbpedia-spotlight.github.com/demo
[3] http://aksw.org/Projects/FOX.html
[4] http://wit.istc.cnr.it/stlab-tools/fred
[5] http://nerd.eurecom.fr
[6] http://viewer.opencalais.com/
[7] http://www.wikimeta.com/wapi/semtag.pl
[8] http://www.zemanta.com/demo/
[9] http://demo.languagecomputer.com/cicerolite

3. Converting Text into RDF Triples

3.1 Overview

A methodology of generating RDF triples from a text in any domain of interest is based on utilization of three tools: AlchemyAPI [8], C&C-Boxer [9] and T2R (Stanford-Senna) [10]. In this chapter, we describe details of the proposed process. A detailed elaboration of each individual step of the procedure is included.

3.2 Process Description

The overview of the proposed methodology is illustrated in Figure 3.1. The input constitutes text paragraphs. At the beginning, the text paragraphs are split into sentences that are pre-processed on a single-sentence basis. Each preprocessed sentence is put into each tool: AlchemyAPI, C&C-Boxer and T2R (Stanford-Senna). Additionally, as seen in Figure 3.1, Stanford Co-reference Resolution System [11] is used to resolve co-references in the triples obtained from C&C-Boxer and T2R tools. All these tools are described in details in the following sections.

In order to obtain appropriate results, the input sentences should be grammatically correct and be in a formal language. For instance, the sentences should not have contraction words like “it’s” instead of "it is". In this work, various stages of creating RDF triples use and analyze different parts of speech of input sentences, such as: proposition phrases, gerunds, and conjunctions. In order to obtain the part of speech information, the output of Senna [14] is utilized due to its capabilities to provide comprehensive information about elements existing in a sentence. Senna is a Natural Language Processing (NLP) tool for studying semantic structure of a sentence. Senna’s output provides part of speech (POS) tag as well as some semantic role labeling for each individual element of a given sentence (see Appendix I for the Senna Sample Output.)

We use Alchemy as an entity recognizer. It has a number of various functionalities. It provides more comprehensive information for each recognized entity which leads to generate

more RDF triples in this study. More information about Alchemy is provided in the section 3.4 of this study.

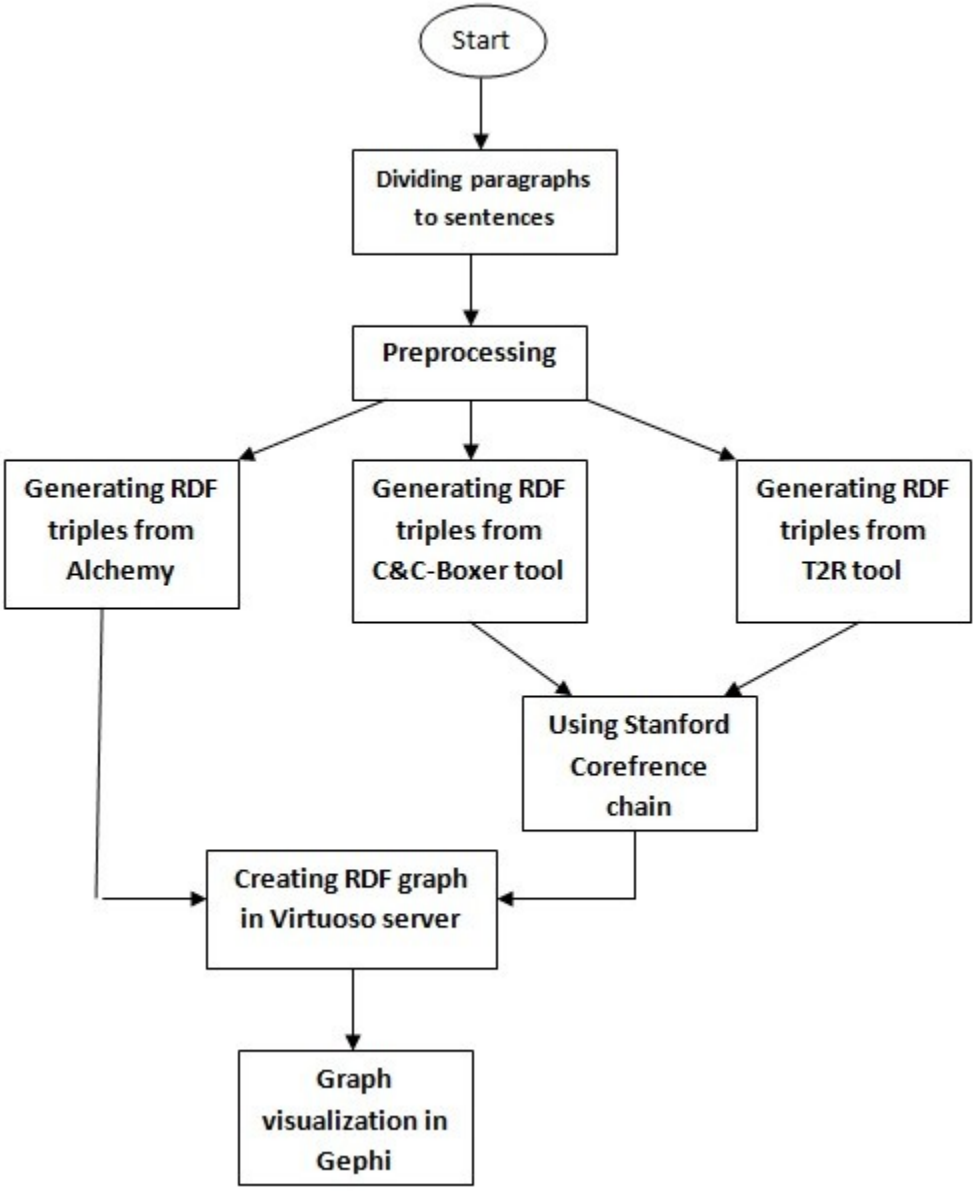

Figure 3-1 the architecture of the study

3.3 Implementation Details

The developed application is called ACT2R (Alchemy, C&C and T2R). It is written in Java programming language while utilizing different libraries. It runs on any Unix-based operating system.

3.4 Pre-processing

A pre-processing stage consists of two processes that are termed: "**Basic P**re-**P**rocessing" (*BasicPP*), and "**Sup**plementary **P**re-**P**rocessing" (*SupPP*). *BasicPP* is applied to each entered sentence. *SupPP*, on the other hand, is applied to complex sentences in addition to the basic preprocessing. Different grammatical structures, which are discussed later, can cause complexity in the sentences. Through experiments conducted in this research, it was observed that in some complex sentences the result of an input processed only by *BasicPP* is less meaningful than the input processed by both *BasicPP* and *SupPP*. Since *SupPP* attempts to divide complex sentences to simple sentences that are easier for parsers to analyze, it has been decided to implement both the basic and the supplementary pre-processing of complex sentences.

However, due to the ambiguities and exceptions observed in experiments, *SupPP* might encounter some problems resulting in incomplete sentences that are difficult to parser. Consequently, in some cases applying both *BasicPP* and *SupPP* entails to meaningless results. To address this drawback, it is proposed that two versions of a sentence are given as inputs to the parsers/tools. The first input is the result of *BasicPP*, while the second input is the result of the sentence processed first via *BasicPP* and then via *SupPP*. The results of parsers are converted to the RDF triples that are further merged. This proposed approach ensure we obtain triples that could be missing if only *BasicPP* or the combination of *BasicPP* and *SupPP* are applied. That is, it can enhance the accuracy and meaningfulness of obtained RDF triples. This approach is adopted due to the fact that recognizing which sentences should be pre-processed with which method is challenging, and would lead to another research activity.

3.4.1 Basic Pre-processing (BasicPP)

Basic Pre-processing (*BasicPP*) performs a set of simple tasks. They include the following activities:

1) Removing comma from digits;
2) Removing any extra symbols such as brackets or parentheses.
3) Removing (an) acronym(s) from sentences and creating (a) triple(s) with property "acronym for": If a sentence has a sequence of words or noun phrases followed by an acronym, each acronym leads to a single triple with the property "acronym for". Afterward, the acronym words are removed from the sentence. The following example illustrates how this pattern is used. In this example, the acronyms "VPH" and "PAHO" are removed after generating the triples with the property "acronym for".

   The analyzed sentence: *"The veterinary public health (VPH) program at the Pan American Health Organization (PAHO) began in 1949."* [22]

   The generated RDF triples are:
   { PAHO, acronym for , Pan American Health Organization . }
   { VPH, acronym for , veterinary public health .}

4) Removing some adverbs such as "in addition", "moreover", "furthermore" in the beginning or middle of the sentences.

3.4.2 Supplementary Pre-processing (SupPP)

The tasks performed in *SupPP* focus on dividing a complex sentence into simple sentences, i.e., sentences that contain in general just a single verb. Such process is very much depending on complexity of a given sentence – its structures, used parts of speech, used propositions and conjunctions.

In order to address this variety and be able to create simple, yet meaningful, sentences we have designed a number of "splitting patterns". These patterns allow for dividing a sentence based on specific components it may contain. The following six patterns are considered in Supplementary Pre-processing (*SupPP*):

1) elimination of a prepositional phrase;
2) elimination of a gerund (i.e. "verb+ing") or past participle form of a verb;
3) division of a sentence containing simple sentences connected via ", and" or "and";
4) division of a sentence containing a proposition and a Relative Pronoun (e.g. "with which");

5) division of a sentence containing a conjunction in the middle of it (e.g. “that”);
6) division of a sentence containing an appositive which is a noun, a noun phrase, or a noun clause which is positioned next to another noun to rename it or to describe it in another way.

Detailed descriptions of each of these patterns are presented below.

1) Elimination of a prepositional phrase: If a sentence starts with a prepositional phrase which can describe location, temporal or manner of the verb of the sentence, then quite T2R (Stanford-Senna) is not able to suitably connect the prepositional phrase to the rest of the sentence. In such a case, the meaning of sentences can be converted into the triples via detecting propositional phrases at the beginning of these sentences, and then connecting them to the rest of the sentences. In addition, the relevant type of adverb of the prepositional phrase is identified via the Senna’s parts of speech (POS) and semantic role labeling outputs. This type is used to build a special triple. In particular, the type, e.g. **where**, **when**, and **how,** becomes a property of the triple. The following example illustrates how this pattern is used.

The analyzed sentence: *“In the southwestern USA, hantavirus was recognized as the cause of a pulmonary syndrome with a mortality rate exceeding 50%.”* [23]

A snippet of the generated RDF triples:

{ hantavirus, recognize as, cause of pulmonary syndrome . }
{ recognize, where, in the southwestern USA . }
{ recognize, patient, hantavirus . }

The Senna’s POS and semantic role labeling are utilized to identify the type of suitable adverbs. In the above example, Senna assigns “AM-LOC” role label to “*southwestern USA*” which indicates that this adverb is the location adverb of the verb “*recognize*”. Therefore, the word **where** is used as a property of the created triple. If Senna identifies “AM-MNR” or “AM-TMP” for the elements of the sentences, the properties of the triples are **how** and **when** respectively.

2) Elimination of a gerund or past participle: If a sentence starts with a gerund (i.e. “verb+ing”) or past participle form of a verb, following by a complete sentence then the utilized

tools (C&C-Boxer and T2R) can barely produce meaningful results. To address this drawback, this process is applied on the sentence which leads to change the sentence's structure without changing its meaning. This helps to obtain a more meaningful and complete result from the parsers, as it is shown in the following instance.

The analyzed sentence: *"Located on the River Thames, London has been a major settlement for two millennia."*

The modified sentence: *"London, located on the River Thames, has been a major settlement for two millennia."*

A snippet of the generated RDF triples:

{ London, type, loc . }

{ London, eq, major settlement . }

{ locate, patient, London . }

{ **locate, on, RiverThames** . }

{ major settlement, for, 2 millennia . }

{ settlement, typeOf, major settlement . }

3) Division of a complex sentence with simple sentences connected via ", and" or "and": If a given sentence includes at least two complete sentences separated with ", and" or "and" conjunctions, its division is required. In some cases, the sentences results from the division are not complete. This happens when their subjects are missing due to the fact that they are contained in the previous parts of a sentence that are new sentences now. These new sentences can also be analyzed using this pattern. Therefore, both of the parsers and AlchemyAPI consider each divided part as a complete sentence. The example of this pattern is as follows:

The analyzed sentence: *"It started as a Section of Veterinary Medicine to help eradicate rabies on both sides of the US-Mexico border, and PAHO grew to be the biggest VPH program in the world."* [22]

Two generated sentences:

"PAHO grew to be the biggest VPH program in the world."

"It started as a Section of Veterinary Medicine to help eradicate rabies on both sides of the US-Mexico border."

The analyzed sentence: *"Member States report the data on animals, feed, food and food-borne outbreaks to EFSAs web-based reporting system and the data on the human cases are reported to ECDCs web - application for The European Surveillance System TESSy ."* [24]

Two generated sentences:

> *"Member States report the data on animals, feed, food and food-borne outbreaks to EFSAs web-based reporting system."*
>
> *"The data on the human cases are reported to ECDCs web application for The European Surveillance System TESSy."*

4) Division of a complex sentence with a preposition and a Relative Pronoun (e.g. "with which"): Having a sentence with a preposition (e.g. "with", "from", and "of") and a relative pronoun (e.g. "which", "whom") consecutively, such as "of which" and "with whom", a complete sentence exists right after the relative pronoun. Therefore, the sentence can be divided to two simple and complete sentences since parses deliver better results with simple sentences. That is, the appropriate triples can be created from the complete sentences. The following example shows how a complex sentence is broken to two sentences.

The analyzed sentence: *"They are characterized by the suddenness, acuteness, the rapidity with which they can spread in susceptible livestock populations and the widespread nature of the losses. "* [25]

Two generated sentences:

> *"They are characterized by the suddenness acuteness, the rapidity."*
>
> *"The rapidity with which they can spread in susceptible livestock populations and the widespread nature of the losses".*

5) Division of a complex sentence with a conjunction existing in the middle of a sentence (e.g. "that"): If a conjunction exists in the middle of a sentence (e.g. "that"), the sentence following the conjunction is either complete or incomplete with a missing subject. In both situations, it is needed to break the sentence at the point where conjunction is located. Additionally, in the case of the incomplete sentence its subject should be found in the previous sentence. The example of this pattern is presented below:

The analyzed sentence: *"The European Community (EC) has been collecting for 15 years data on zoonoses and agents that integrate the information from human cases and their occurrence in food and animals."* [24]

Two generated sentences:

> *"The European Community (EC) has been collecting for 15 years data on zoonoses and agents"*
>
> *" zoonoses and agents integrate the information from human cases and their occurrence in food and animals."*

6) Construction of two simple sentences if a sentence has an appositive which is a noun, a noun phrase, or a noun clause which is positioned next to another noun to rename it or to describe it in another way: If a sentence has an appositive which is a noun, a noun phrase, or a noun clause which is located next to another noun with the purpose of renaming it or describing it in another way, then there is a need to create another sentence. This new sentence is made from the appositive noun or noun phrase. The process ensures that generated triples cover the essence of the sentence. The following example shows how this pattern applies in the sentence.

The analyzed sentence: *" London, the capital of England and the United Kingdom, was founded 2000 years ago by the Romans as Londinium."*

Two generated sentences:

> *"London was founded 2000 years ago by the Romans as Londinium."*
>
> *"London is the capital of England and the United Kingdom"*

After applying all pre-processing steps, all sentences become an input to Alchemy, C&C-Boxer, and T2R (Stanford-Senna) tools.

3.5 Generating RDF Triples from AlchemyAPI

In this study, AlchemyAPI [8] is used as an entity recognizer. In the following sections, first, a brief description of AlchemyAPI is provided. The procedure of creating RDF triples from this tool is, then, explained.

3.5.1 AlchemyAPI

AlchemyAPI is a web service with a powerful natural language processing technology which analyzes any type of text and identifies named entities such as people, locations, organizations, as well as facts and relations, topic keywords, text sentiment, news and blog article authors, taxonomy classifications, scraping structured data, and more. [8].

AlchemyAPI can be used directly through either the internet-interface or its downloadable software developer kits (SDKs). SDKs are provided in various programming languages including Java, C/C++, C#, Perl, PHP, Python, Ruby, JavaScript and Android OS. AlchemyAPI is able to process even small utterances like Twitter posts [8]. However, a longer text is more preferable as an input of API [8].

In order to have full access to AlchemyAPI, it is needed to obtain an access key which should be considered confidential and not shared with anyone or embedded within any package (software, etc) that is publicly distributed. [8].

3.5.2 AlchemyAPI Entity Extraction

AlchemyAPI can recognize people, companies, organizations, cities, geographic features, and other typed entities within HTML, text, or web-based contents. To this end, advanced statistical algorithms and natural language processing technology are implemented to analyze information, and extract the semantics of the content [8]. Information on different types of entities is included in [8]. AlchemyAPI is featured by unique combination of advanced multi-lingual support, RDF/Linked Data, context-sensitive entity disambiguation, comprehensive type support, and quotations extraction [8].

Different formats can be utilized for extracted meta-data such as XML, JSON, RDF, and Microformats rel-tag formats [8]. In this study, XML output format is used. Once the output is obtained, it is utilized to generate RDF triples.

3.5.3 AlchemyAPI Entity Disambiguation

In order to resolve a company, location, or individual into a unique instance, AlchemyAPI utilizes a sophisticated “entity disambiguation” since some entities need the surrounding context to recognize the real type of the entities and relevant information [8]. Therefore, the surrounding

context can be used to resolve the ambiguities of the named entities. For instance, when disambiguating an individual, the information such as the person's location, profession or employer can be used. As such, for a company, the information on the company's key executives, notable products, industry, location, etc are use [26]. For further information on AlchemyAPI Entity Disambiguation mechanism see [26]

3.5.4 RDF Triples from AlchemyAPI

The following example shows the XML output of named entities recognized by AlchemyAPI for a given sentence.

The analyzed sentence: *"The European Food Safety Authority (EFSA) is assigned the tasks of examining the data collected and publishing the Community Summary Report."* [24]

XML output of Alchemy

<?xml version="1.0" encoding="UTF-8" standalone="no"?><results>
<status>OK</status>
<usage>By accessing AlchemyAPI or using information generated by AlchemyAPI, you are agreeing to be bound by the AlchemyAPI Terms of Use: http://www.alchemyapi.com/company/terms.html</usage>
<url/>
<language>english</language>
<entities>
<entity>
<**type**>Organization</**type**>
<relevance>0.33</relevance>
<count>2</count>
<**text**>EFSA</**text**>
<disambiguated>
<name>European Food Safety Authority</name>
<website>http://www.efsa.europa.eu/</website>
<dbpedia>http://dbpedia.org/resource/European_Food_Safety_Authority</dbpedia>
<freebase>http://rdf.freebase.com/ns/m.094yn1</freebase>
<yago>http://yago-knowledge.org/resource/European_Food_Safety_Authority</yago>
</disambiguated>
</entity>
</entities>
</results>

This XML output of Alchemy for Named Entities is used to generate RDF triples as shown below:

RDF Triples generated based on Alchemy's output:

{ EFSA, website, http://www.efsa.europa.eu/ . }

{ EFSA, SameAs, http://dbpedia.org/resource/European_Food_Safety_Authority . }

{ EFSA, SameAs, http://rdf.freebase.com/ns/m.094yn1 . }

{ EFSA, SameAs, http://yago-knowledge.org/resource/European_Food_Safety_Authority . }

{ EFSA, Type, Organization . }

As we can see in the XML output the "text" tag defines the name of the entity. Another important tag is "type" that is used directly to create a new RDF triple with the property "Type" (i.e. "EFSA, Type, Organization"). Then, if there are any "disambiguated" tags, as seen in the example, a number of RDF triples are created from them as explained in the following description.

As we could observe in the previous paragraph, the names of the XML tags are used to create the properties of generated RDF triples, i.e, the tag **type** leads to the property **type**, the XML tag **disambiguated** leads to the property **SameAs**. The value of the XML tag **text** is the subject of the RDFs, and the content of the **disambiguated** tag as the object of the triples with the property **SameAs**.

One of the benefits of the Alchemy is its ability to provide references to relevant web pages of each named-entity. These web pages provide more elaborated information about the named entities. They are such pages such as dbpedia, freebase, yago, etc.

In the presented study, the Alchemy entity recognizer feature plays the role of a reference. It means that if the type of an entity is recognized as a different category using other tools, i.e., C&C-Boxer, T2R, the type assigned to the entity is the one obtained from Alchemy. For instance, some entities in C&C-Boxer or T2R that are recognized as a "location" type are of the type "organization" as Alchemy API indicates, and Alchemy does it correctly.

3.6 Generating RDF Triples from C&C-Boxer System

In this study, C&C-Boxer system is used as one of the sources of RDF triples. In the following sections, we briefly describe C&C-Boxer, and then we explain the procedure used for creating RDF triples from the output of C&C-Boxer.

3.6.1 C&C-Boxer System

C&C-Boxer is a Natural Language Processing (NLP) system using syntactic and semantic forms created based on theoretical linguistics [9]. C&C, which is a statistical parser, is created based on Combinatory Categorical Grammar (CCG) parser [27]. Using C&C-Boxer system, relationships between entities are identified. Initially, the sentence's words are tagged as parts of speech using the Penn Treebank tagset (see Appendix II for the list of POS tags used in the Penn Treebank). C&C also includes a named entity recognizer that distinguishes between ten different entity types with the following labels [15]: Org (organization) ,Per (person), Ttl (title), Quo (quotation), Loc (location), Fst (first name), Sur (surname), Url (URL),Ema (e-mail), and Nam (unknown name).

The Boxer part has been developed by Johan Bos [28]. It is a separate component that uses the output of C&C parser. It performs a deep semantic analysis and produces interpretable structure in the form of Discourse Representation Structures (DRSs) [9]. Curran et el. [9] has presented the following description on DRSs:

"DRSs are recursive data structures—each DRS comprises a domain (a set of discourse referents) and a set of conditions (possibly introducing new DRSs). DRS-conditions are either basic or complex. The basic DRS-conditions supported by Boxer are: equality, stating that two discourse referents refer to the same entity; one-place relations, expressing properties of discourse referents; two place relations, expressing binary relations between discourse referents; and names and time expressions. Complex DRS-conditions are: negation of a DRS; disjunction of two DRSs; implication (one DRS implying another); and propositional, relating a discourse referent to a DRS."

In this description, the term "discourse referents (DR)" refers to relevant entities, and the term "conditions" are relationships between the discourse referents [15].

Discourse referents are defined using new noun phrases or events [15]. For every relation existing between discourse referents, a condition is created. It can be a unary predicate (indicating unary relation) introduced by nouns, verbs, adverbs and adjectives, or it can be a binary predicate (indicating binary relation) introduced by propositions and verb roles, e.g., agent, patient, or theme [15]. It should be emphasized that since DRs conditions only have unary and binary relations, their structures are similar to RDF structure. Thus, the output of the Boxer provides a suitable structure to be used for converting text into RDF triples [15].

The C&C tools and Boxer are accessible for downloading from the following website [29].

3.6.2 Semantic of Boxer System

In order to create triples from Boxer outputs which is based on the output of C&C parser, it is required to translate the semantic of the Boxer in a way to achieve to desired RDF triples that should be simple and meaningful.

The following example shows the output of Boxer for the sentence given below. The C&C output is included in Appendix III.

The analyzed sentence: *"Another visible accomplishment is the elimination of hydatidosis in the endemic countries and regions of the southern cone."* [22]

The Boxer output:

```
sem(1,
   [
    word(1001, 'Another'),
    word(1002, visible),
    word(1003, accomplishment),
    word(1004, is),
    word(1005, the),
    word(1006, elimination),
    word(1007, of),
    word(1008, hydatidosis),
    word(1009, in),
    word(1010, the),
    word(1011, endemic),
    word(1012, countries),
```

```
 word(1013, and),
 word(1014, regions),
 word(1015, of),
 word(1016, the),
 word(1017, southern),
 word(1018, cone),
 word(1019, '.')
],
[
 pos(1001, 'DT'),
 pos(1002, 'JJ'),
 pos(1003, 'NN'),
 pos(1004, 'VBZ'),
 pos(1005, 'DT'),
 pos(1006, 'NN'),
 pos(1007, 'IN'),
 pos(1008, 'NN'),
 pos(1009, 'IN'),
 pos(1010, 'DT'),
 pos(1011, 'JJ'),
 pos(1012, 'NNS'),
 pos(1013, 'CC'),
 pos(1014, 'NNS'),
 pos(1015, 'IN'),
 pos(1016, 'DT'),
 pos(1017, 'JJ'),
 pos(1018, 'NN'),
 pos(1019, '.')
],
[
],
[
 l0:alfa(def, l1, l4),
 l1:drs([[]:_G455], [l2, l3]),
 l2:[1002]:pred(_G455, visible, a, 0),
 l3:[1003]:pred(_G455, accomplishment, n, 0),
 l4:merge(l5, l11),
 l5:drs([[1001]:_G489], [l6, l7, l8]),
 l6:[1002]:pred(_G489, visible, a, 0),
 l7:[1003]:pred(_G489, accomplishment, n, 0),
```

```
 l8:[1001]:not(l9),
 l9:drs([], [l10]),
 l10:[]:eq(_G489, _G455),
 l11:alfa(def, l12, l36),
 l12:merge(l13, l15),
 l13:drs([[1005]:_G547], [l14]),
 l14:[1006]:pred(_G547, elimination, n, 0),
 l15:merge(l16, l34),
 l16:merge(l17, l19),
 l17:drs([[1008, 1009, 1010, 1011, 1012, 1013, 1014, 1015, 1016, 1017, 1018]:_G573], [l18]),
 l18:[1008]:pred(_G573, hydatidosis, n, 0),
 l19:alfa(def, l20, l27),
 l20:drs([[1010]:_G597], [l21, l22]),
 l21:[1011]:pred(_G597, endemic, a, 0),
 l22:[1013]:or(l23, l25),
 l25:drs([], [l26]),
 l26:[1014]:pred(_G597, region, n, 0),
 l23:drs([], [l24]),
 l24:[1012]:pred(_G597, country, n, 0),
 l27:alfa(def, l28, l31),
 l28:drs([[1016]:_G658], [l29, l30]),
 l29:[1017]:pred(_G658, southern, a, 0),
 l30:[1018]:pred(_G658, cone, n, 0),
 l31:drs([], [l32, l33]),
 l32:[1015]:rel(_G597, _G658, of, 0),
 l33:[1009]:rel(_G573, _G597, in, 0),
 l34:drs([], [l35]),
 l35:[1007]:rel(_G547, _G573, of, 0),
 l36:drs([[1004]:_G728], [l37, l38]),
 l37:[]:pred(_G728, event, n, 1),
 l38:[1004]:prop(_G728, l39),
 l39:drs([], [l40]),
 l40:[1004]:eq(_G489, _G547)
] ).
```

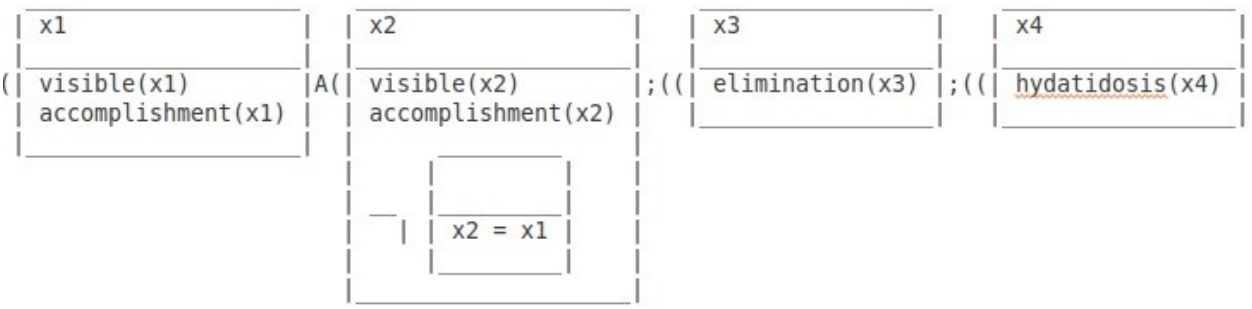

Figure 3-2 Part of Box-like structure of Boxer

As seen in the above example, an ID with the label “word” is assigned to each word in the sentence, for example “word (1002, visible)”. Then, parts of speech (POS) for those words are determined. It should be noted that in some cases, C&C-Boxer might recognize Named Entities and label them with “ne” and their corresponding types. The following example shows the named Entities recognized in the sentence below. The ID assigned to each Named Entity refers to the ID of corresponding word in the sentence. For example ID 1001 is assigned to the word “London” as well as to the Named Entity 'I-LOC' .

The analyzed sentence: *“London is the capital of England and the United Kingdom.“*

Named Entities recognized by C&C-Boxer tool:

ne(1001, 'I-LOC'),
ne(1006, 'I-LOC'),
ne(1009, 'I-ORG'),
ne(1010, 'I-ORG')

The last section of the Boxer’s output, called DRS hereafter, is converted into RDF triples using the following process (Section 3.6.3). Ultimately, the Boxer output is presented in a box-like structure, Figure 3.2.

3.6.3 Procedure of Creating RDF Triples from C&C-Boxer Output

To create RDF triples based on the output generated by Boxer, a procedure consisting of a number of tasks and required modifications is designed. These tasks, which make the converting process for generating meaningful triples simple and doable, are stated below.

The first task is to store all words from a sentence with the IDs assigned to them as a two-tuple set (see"Two-tuple set" below). All words with the same ID are kept in the same tuple as seen in the following example.

Two-tuple set:

G455, visible accomplishment
G489, visible accomplishment
G547, elimination
G573, hydatidosis
G597, endemic countries regions
G658, southern cone
G728, event

As seen in this example, the words "visible" and "accomplishment" have the same ID (i.e. G455) and they are stored in the same two-tuple as "G455, visible accomplishment".

The next task is to find all relationships between the words defined/used in the sentence. Once they are found they are stored together with the words and their IDs in the form of a three-tuple set (see "Three-tuple set" below). The following is the example of such a task.

Three-tuple set**:**

G489, G455, eq
G597, G658, of
G573, G597, in
G547, G573, of
G489, G547, eq

Matching the word to IDs, the following relationships are extracted from the presented triples.

visible accomplishment	,	eq	,	elimination
elimination	,	of	,	hydatidosis
hydatidosis	,	in	,	endemic countries regions
endemic countries regions	,	of	,	southern cone

It should be noted that the label “eq” refer to “equals” which is eventually replaced with the related to be verb such as “is” and “are” in the final RDF triples.

There are also a number of tasks that are preformed to make the generated triples simpler and more descriptive. All of these tasks are listed and described below. A description of each activity finishes with a simple and illustrative example.

Restoring plural forms of nouns: The Boxer’s output contains the basic format of each verb or noun. Therefore, in order to have results that can be compared and matched to the results obtained from T2R, it is needed to find out the origin of the words as they are in input sentences. The words in singular format from the Boxer’s output are compared with the words in the sentences, and then they are replaced by their plural format. For example, in the previously analyzed sentence *“Another visible accomplishment is the elimination of hydatidosis in the endemic countries and regions of the southern cone”* the word “country” is found in the Boxer’s output and is replaced with “countries”.

Restoring capital letters at the beginning of words: In the case of words that start with capital letters they are converted to the lower case in the Boxer’s output. They are replaced by their origin words as shown in the following example.

The analyzed sentence: *“The **Pan American** centers developed a number of diagnostic antigens.”* [22]

A fragment of Boxer’s output:

l4:[1002, 1003]:named(_G12470, **pan_american**, org, 0),

A portion of the generated RDF triples:

{ **PanAmerican** centers, typeOf, centers . }

{ **PanAmerican**, type, org . }

Restoring alphabetic format of numbers: Boxer converts numbers from an alphabetic format to numerical one. Therefore, it is required to convert numbers back from the numerical format to the original verbal format as shown in the following example.

The analyzed sentence: *"The American Society for Microbiology is **one** of many partners in the LRN."* [30]

A fragment of Boxer's output:

l12:[1007]:card(_G17828, **1**, ge),

A portion of the generated RDF triples:

{ AmericanSociety, for, Microbiology . }

{ **one** of partners, in, LRN . }

{ AmericanSociety, eq, **one** of partners . }

{ **one** of partners, typeOf, partners . }

{ **one** of partners, typeOf, **one** . }

Removing extra symbols from words: In the Boxer's output, it is observed that any sequence of words starting with capital letters exists in a given sentence is converted into a group of words separated with hyphens. Such a construct is not desirable, and the words are converted back to their original format. It is shown in the following example.

The analyzed sentence: *"The Pan American centers developed a number of diagnostic antigens."* [22]

A fragment of Boxer's output:

l4:[1002, 1003]:named(_G12470, **pan_american**, org, 0),

A relevant triple:

{ **PanAmerican** centers, typeOf, centers . }

Creating triples with the property *"typeOf"*: One of the relationships that Boxer can identify is the relationship "*nn*" between words in noun phrases. Since one of the goals of this work is to create simple triples it means that all elements of a single triple (i.e. subject, property and object) should be simple. Therefore, the proposed approach generates a number of triples based on the "*nn*" property. This is accomplished via creating triples with the property "*typeOf*".

In this case, when there is a triple with nodes containing more than one noun, a number of triples that contain the origin or root of these nodes are created. The example of such a scenario is presented below.

The analyzed sentence: *"This network of infectious diseases consultants was conceived as a sentinel system to monitor new or resurgent infectious diseases in a way that would complement other public health surveillance efforts."* [31]

Two-tuple set:

G644, disease

G650, network

G656, infectious consultant

G1055, surveillance

G1061, health

G955, public efforts

Three-tuple set:

G644, G656, nn

G650, G656, of

G1055, G955, nn

G1061, G955, nn

The triples created based on the relationships defined in the Three-tuple set:

{ diseases, typeOf, diseases infectious consultants . }

{ surveillance, typeOf, surveillance public efforts . }

It should be noted that the order of nouns in the triples created based on the "*nn*" relationships identified by the Boxer is not always correct – this can be seen in the generated triples above. In order to achieve more accurate results, a location of the each element in the sentence should be identified. This information is used to make the order of elements of a given phrase correct. To achieve this, the "*word*" and "*pos*" entries of the Boxer's output are utilized. When a combined

word (phrase) exists in the created triples, the correct order is obtained using the POS (part of speech) of the words, as well as their locations in the sentence. The following example shows how this approach changes the results.

Triples created without the proposed approach:

{ diseases, typeOf, diseases infectious consultants . }

{ surveillance, typeOf, surveillance public efforts . }

Triples created with the proposed approach:

{ diseases, typeOf, infectious diseases consultants . }

{ surveillance, typeOf, public surveillance efforts . }

Creating triples with verbs' arguments: A very interesting relationship identified by Boxer is related to a link between a verb and its arguments. Let us illustrate such a situation with an example.

The analyzed sentence: *"The European Community (EC) has been collecting for 15 years data."* [24]

A fragment of Boxer's output

l2:[1002, 1003, 1004]:named(_G3184, 'european_community', org, 0),

l13:[1011]:pred(_G3415, datum, n, 0),

l19:[1007]:pred(_G3520, collect, v, 0),

l23:[1007]:rel(_G3520, _G3092, agent, 0),

l24:[1007]:rel(_G3520, _G3323, patient, 0),

l25:[1008]:rel(_G3520, _G3478, for, 0),

l11:[1009]:card(_G3478, 15, ge),

l16:[1010]:pred(_G3478, year, n, 0),

A part of theTwo-tuple set:

G3415, datum

G3520, collect

G3184, european_community

G3478, 15 year

A part of the Three-tuple set:

G3520, G3184, agent

G3520, G3415, patient

G3520, G3478, for

Based on the relationships defined in the **Three-tuple set,** the following triples are created.

{ collect, agent, EuropeanCommunity . }

{ collect, patient, data . }

{ collect, for, 15 years . }

The two of above triples: the triple with the property "*agent*", and the triple with the property "*patient*", are converted into more meaningful triples of the following format:

{ EuropeanCommunity, collect, data . }

{ collect, for, 15 years . }

It is seen that those two triples (with the properties "*agent*" and "*patient*") are converted into a unique triple that contains the subject – "EuropeanCommunity" – from the first triple, the property – "collect" – that is the subject of all triples, and the object – "data" – that is the object of the second triple.

It should be emphasized that the tasks described previously, such as "**Restoring plural forms of nouns**" and "**Removing extra symbols from words**" are also performed to convert "datum" to the plural format ("data") and removing the hyphen from "'european_community".

Creating triples based on verbs connected with proposition: In some cases, a sentence contains verbs connected with "to" proposition. Such a construct requires a special post-processing of triples that are initially generated from the sentence. Let us take a look at the following example.

The analyzed sentence: *"... other activities are designed to benefit consultants in infectious diseases."* [31]

Initially generated triples:

{ design, agent, activities . }

{ design, theme, proposition . }

{ design, theme, proposition . }

{ activities, benefit, consultants . }

The above triples are converted to a more meaningful triple:

{ activities, design to benefit, consultants . }

Creating triples from sentences with negative meaning: There are two forms of sentences that have negative meaning: sentences with negative verbs, and sentences with negative noun phrases. The Boxer uses DRS (Discourse Representation Structures)–condition, shown in bold in the Boxer outputs for given sentences presented in Example 1 and Example 2, to express a negation. It is complex and it is not easy to determine which part of a given sentence has been negated. Therefore, when a negative syntax is detected in the Boxer's output (see the examples below), it is necessary to identify any negative word in the sentence and find which elements of the sentence have been negated. To accomplish this, we look for any negated word in the sentence. These negative words do not necessarily negate the word located exactly before or after itself. Since Boxer's output keeps only the main verb of any tense in the sentence, it is required to consider all different kinds of tense and grammar structures with or without an adverb to identify what would be the main negated element (verb or noun). In order to create triples, it is needed to find those negated words and place negations before them. The followings examples with different grammar structures demonstrate how this approach works.

Example 1:

The analyzed sentence: *"As TADs do not recognize national borders."* [25]

A fragment of Boxer's output:

l0:drs([], [l1]),

l1:[1003]:not(l2),

l2:drs([[1001]:_G8354, [1005, 1006]:_G8375, [1004]:_G8396], [l3, l4, l5, l6, l7, l8, l9]),

l3:[1001]:pred(_G8354, tad, n, 0),

l4:[1005]:pred(_G8375, national, a, 0),

l5:[1006]:pred(_G8375, borders, n, 0),

l6:[1004]:pred(_G8396, recognize, v, 0),

l7:[]:pred(_G8396, event, n, 1),

l8:[1004]:rel(_G8396, _G8354, agent, 0),

l9:[1004]:rel(_G8396, _G8375, patient, 0)

A portion of the generated triples:

{ **not recognize**, agent, TADs . }

{ **not recognize**, patient, national borders . }

{ TADs, **not recognize**, national borders . }

{ national borders, typeOf, borders . }

Example 2:

The analyzed sentence: *"... the end of the epidemic is not yet in sight."* [32]

A fragment of Boxer's output

l0:drs([], [l1]),

l1:[1007]:not(l2),

l2:alfa(def, l3, l11),

l3:merge(l4, l6),

l4:drs([[1001]:_G14120], [l5]),

l5:[1002]:pred(_G14120, end, n, 0),

l6:alfa(def, l7, l9),

l7:drs([[1004]:_G14177], [l8]),

l8:[1005]:pred(_G14177, epidemic, n, 0),

l9:drs([], [l10]),

l10:[1003]:rel(_G14120, _G14177, of, 0),

l11:drs([[1006]:_G14259, [1010]:_G14280], [l12, l13, l14, l15, l16, l17]),

l12:[1006]:pred(_G14259, be, v, 0),

A portion of the generated triples:

{ **is not yet**, in, sight . }

{ **is not yet**, typeOf, is not . }

{ **is not yet**, agent, end of epidemic . }

{ end of epidemic, typeOf, epidemic . }

{ end of epidemic, typeOf, end . }

Example3:

The analyzed sentence: *"We cannot be complacent about our contingency planning."* [33]

A portion of the generated triples:

{ event, about, contingency planning . }

{ person, **not be**, complacent . }

{ contingency planning, typeOf, planning . }

In the triples from Example3, the word "person" that refers to the pronoun "we" is eventually replaced with an appropriate word – this is resolved via application of the Stanford Corefrence chain described later.

Creating triples explaining adverbs included in a sentence: Other important and interesting parts of a sentence that are required to make generated RDF triples more comprehensive are the elements that are adverbs describing locations of actions, their temporal aspects, as well as manners or degrees of actions. These types of adverbs are marked with specific labels used by "Senna": "AM_LOC" for location adverbs, "AM_TMP" for temporal adverbs, and "AM_MNR" for manner adverbs. For this purpose, the Senna's POS and semantic role labeling output are explored to find the above-mentioned labels. Additionally, corresponding verbs that these adverbs describe are discovered. Afterwards, a new triple is created with the verb as a "subject" and with an appropriate property – "where" for a location adverb, "when" for any temporal adverb, or "how" for any "manner" adverb. The following example illustrates utilization of this pattern.

The analyzed sentence: *"More than 300 infectious diseases consultants are currently participating in the IDSA EIN."* [31]

A portion of the generated triples:

{ IDSAEIN, type, loc . }

{ participate, in, IDSAEIN . }

{ **participate, when, currently .** }

For the above sentence, Senna assigns the label “AM_TMP” to the word “currently” which leads to have a triple with the property “when” as shown in the above triples.

3.7 Generating RDF triples from T2R (Stanford-Senna) Tool

In this work, we also use another system called T2R (Stanford-Senna) [10] to create meaningful and simple RDF triples. T2R applies two methods to process every sentence: *Stanford Parser* and *Senna*. *Stanford Parser* [13], is a statistical parser of natural language that works with the grammatical structure of sentences of a given raw text. *Senna* Parser [14] is the semantic parser capable of predicting: part-of-speech (POS) tags, chunking (CHK), name entity recognition (NER), semantic role labeling (SRL) and syntactic parsing (PSG).

The output of T2R is a set of RDF triples. However, we further process these RDF triples to generate triples that are more meaningful and simpler. The following example shows triples generated by T2R.

The analyzed sentence: *“There exists a huge growth area for the veterinary profession.”* [34]

T2R triples:

area-6 is thing-existing
area-6 for profession-10
area-6 exists-direct-object area-6
profession-10 typeOf profession_veterinary
area-6 typeOf huge_growth_area

A number of modifications performed on the triples generated by T2R are listed and explained.

Removing numbers and symbols: First, as seen in the above example, each node is associated with a number, e.g. “6” is assigned to “area”. This number refers to the location of a word in a given sentence. These numbers are removed. Additionally, an extra symbol “_” is used to separate words in combined noun phrases, e.g. “_” in “huge_growth_area ”. This is also removed.

Recovering basic formats of verbs: T2R keeps the original format of the verb as it is used in a sentence. As we can see in the above example, the verb "exists" in the node "exists-direct-object". It is preferable to have the basic format of the verb. This would make the triples generated by T2R comparable to the Boxer's triples. Therefore, any verb is replaced by its basic format. This is accomplished using the verb recognizing functions from two Java libraries:

- "edu.smu.tspell.wordnet.WordNetDatabase", and

- "edu.smu.tspell.wordnet.impl.file.Morphology".

These functions sometime provide several basic formats of a verb, and only one of them is correct. The obtained basic formats are compared with corresponding basic verbs provided by Boxer in order to identify the correct verb. This identified verb is used in the triples. For instance, the Java functions provide "goe" and "go" as the basic format of the verb "going". Both of them are compared with the corresponding output of Boxer to identify the correct verb: "go".

Generation of triples with PropBank roles: The goal of creating meaningful RDF triples from the T2R output is supported by utilization of PropBank [18]. The PropBank is a set of verbal propositions and numbered arguments (e.g.*Arg0, Arg1, etc*.). Since some triples generated by T2R system have nodes indicating different arguments of verbs, new RDF triples can be created based on them. PropBank recognizes a number of verb arguments that are listed in Table 3.1:

Table 3.1 List of Args recognized by PropBank

Numbered Argument	**Description of the arguments**
Arg0	an agent, causer who does an experiment
Arg1	theme, patient
Arg2	instrument, benefactive, attribute
Arg3	starting-point
Arg4	ending-point
ArgM	describe different modifiers determined by a context, they can be Temporal, Locative, and Directional modifiers

The following example shows T2R triples generated for a given sentence together with PropBank roles of the verb in that sentence.

The analyzed sentence: *"A pilot study with 169 participants recruited from 32 of the IDSAs state and regional societies confirmed the feasibility and potential value of this network."* [31]

T2R Triples:

participants-6 typeOf 169_participants
32-9 of state-13
study-3 is recruit
study-3 with participants-6
IDSAs-12 is MISC
societies-16 typeOf regional_societies
study-3 recruited-from 32-9
32-9 of societies-16
32-9 is group
state-13 typeOf IDSAs_state
study-3 typeOf pilot_study
value-22 is thing-confirmed
participants-6 typeOf 169_participants
value-22 of network-25
value-22 typeOf potential_value_feasibility
study-3 with participants-6
study-3 is confirmer
study-3 confirmed-direct-object value-22
study-3 typeOf pilot_study

PropBank roles:

confirmed { 'A1': 'thing-confirmed,-statement-or-proposition'

'A0': 'confirmer',

'A2': 'benefactive,-hearer'}

Using the PropBank roles defined for the verb "confirmed" and finding those arguments among the objects of T2R triples, a new triple consisting of the following elements is created:

Subject: "study" is the subject of the triple with the object "confirmer" (Arg0 defined in PropBank);

Property: "confirm" the corresponding verb;

Object: "value" is the subject of the triple with the object "thing-confirmed" (Arg1 defined in PropBank).

Therefore, the following triple is created:

{ study, confirm, value }

As seen in the above example, values of different dependencies or arguments – the word "study" as the subject, and the word "value" as the object of the verb "confirm" – can be found in other triples generated by T2R. The T2R triples with the "typeOf" property show the origin or details of each complex node. In this case, not only do we make our triple more meaningful but also we keep the simplicity of the nodes by having the triples with "typeOf" property. Hence, the complete value of PropBank arguments as well as the complete triple can be created as follows:

Complete value of PropBank arguments of the verb "confirm"

potential value feasibility, is ,Thing-confirmed

pilot study, is ,Confirmer

Complete triple created using PropBank arguments:

{**pilot study**, confirm, **potential value feasibility of network** }

In some situations, a sentence is in the form of the passive voice or T2R is not able to determine the value of Arg0 of a verb. In such cases, the Arg0 of the PropBank role does not have any value. Hence, the subject of the new triple is replaced with the name of the Arg0 itself. The following examples show the application of this pattern.

Example 1:

The analyzed sentence: *"The current data collection covers 11 zoonotic agents."* [24]

T2R triples:

collection-4 covers-direct-object agents-8

collection-4 typeOf current_data_collection

collection-4 is instrument

agents-8 is thing-covered

agents-8 typeOf zoonotic_agents

agents-8 number 11-6

PropBank roles:

covers { 'A1': 'thing-covered',
'A0': 'coverer',
'A2': 'instrument,-covered-by' }

The modified triples:

{ current data collection, cover, 11 agents . }

{ current data collection, typeOf, collection . }

{ current data collection, is, Instrument . }

{ 11 agents, is, Thing-covered . }

{ zoonotic agents, typeOf, agents . }

{ coverer, cover, 11 agents . }

{ 11 agents, typeOf, agents . }

As seen in Example 1, there are no T2R triples with the value of Arg0 as defined in PropBank for the verb "covers". Therefore, the name of Arg0 – "coverer" – is considered as the subject of a new triple with property "cover" – "{coverer, cover, 11 agents . }" .

Example 2:

The analyzed sentence: *"The flow and analysis of data are described."* [24]

T2R triples:

flow-2 of data-6

flow-2 is thing-described

analysis-4 is thing-described

PropBank roles:

described { 'A1': 'thing-described',
'A0': 'describer',
'A2': 'secondary-attribute,-described-as'}

The modified triples:

{ describer, described, flow of data . }

{ describer, described, analysis. }

Also in this example, a value of Arg0 of the verb ‘described” is not identified. Therefore, the name of Arg0, i.e. “describer”, is used as a subject of the triple produced based on the PropBank roles: “{describer, described, flow of data.}”, and “{describer, described, analysis.}”.

Generation of more triples with “typeOf” property: for some combined words or noun phrases consisting conjunctions like “of”, T2R creates triples of which properties are the conjunctions (e.g. “32-9 of societies-16” in the following example). We can convert those triples to triples with “typeOf” properties as exhibited in the following example:

The analyzed sentence: *“A pilot study with 169 participants recruited from 32 of the IDSA's state and regional societies confirmed the feasibility and potential value of this network.”* [31]

A portion of T2R triples:

societies-16 typeOf regional_societies

study-3 recruited-from 32-9

32-9 of societies-16

study-3 recruited-from 32-9

study-3 typeOf pilot_study

A portion of the modified T2R triples:

{ pilot study, recruit from, **32 of regional societies** . }

{ pilot study, typeOf, study . }

{ 32 of regional societies, typeOf, societies . }

{ 32 of regional societies, typeOf, regional societies . }

{ regional societies , typeOf ,societies}

As seen in the above example, the object of the triple with the property “recruited-from” in the T2R triples is replaced by the complete phrase with “of” conjunction in the modified triple.

In addition, it can be also observed that the place of subject and object of any “typeOf” triples are exchanged to make these triples more meaningful. It should be emphasized that this modification is generally applied to any triple with the “typeOf” property in triples generated by T2R (it is explained in more detail later).

Modification of triples with the property "number": If a number exists before a noun in a given sentence, T2R produces a triple with "number" property. In such a case, we find all the triples that have the same subject as the triple with the "number" property, and then we append the object of this triple to the subject. It is illustrated in the following example.

The analyzed sentence: *"More than 300 infectious diseases consultants are currently participating in the IDSA EIN."* [31]

T2R triples:

consultants-6 **number** 300-3
IDSAEIN-12 is LOC
currently-8 is AM-TMP
consultants-6 is agent
consultants-6 participating-in IDSAEIN-12
consultants-6 typeOf infectious_diseases_consultants
IDSAEIN-12 is participating-in-what?
consultants-6 participating-more-detail currently-8

The modified T2R triples:

{ IDSAEIN, is, Loc . }
{ **300 consultants**, is, Agent . }
{ **300 consultants**, participate in, IDSAEIN . }
{ infectious diseases consultants, typeOf, consultants . }
{ IDSAEIN, is, Participating-in-what? . }
{ participate, when, currently . }
{ **300 consultants**, participate, IDSAEIN . }
{ **300 consultants, typeOf, consultants** . }

As seen in the above modified triples, other triples, for example "{ 300 consultants, typeOf, consultants . }" that consist of the following nodes have to be created to keep the simplicity of the results.

Subject: appended subject

Property: "typeOf"

Object: object of the triple with the property "number"

Modification of triples with the property "more-detail" for verbs: T2R is able to generate triples with "more-detail" property as it can be seen in the following example:

The analyzed sentence: *"it is localized and then a disease control program be quickly implemented."* [25]

T2R triples:

program-9 typeOf disease_control_program

implemented-12 more-detail quickly-11

localized-3 subject it-1

implemented-12 subject program-9

quickly-11 is AM-MNR

As seen in the above example, the "more-detail" property provides some extra information about the adverb describing the verb. In order to identify the type of adverb, the Senna's semantic role labeling are utilized. It provides the label "AM-MNR" marking the word "quickly" and this reveals that "quickly" is a manner adverb. In this case, the property "more-detail" is replaced with the property "how" as follows:

The modified triple:

{ implement , how, quickly . }

If the Senna's semantic role labeling assigns "AM-LOC" and "AM-TMP" labels to the elements of sentences, the corresponding properties of triples are "where" and "when" respectively.

Modification of triples with the property "verb+more-detail": In some cases, the property of a triple contains a "verb + more-detail" label. In these situations, we utilize the Senna's semantic role labeling and perform the procedure similar to the one from the previous case:

another triple with an appropriate property such as “where”, “where”, “how” is created. It is shown in the following example.

The analyzed sentence: *“More than 300 infectious diseases consultants are currently participating in the IDSA EIN.”* [31]

T2R triples:

consultants-6 number 300-3

IDSAEIN-12 is LOC

currently-8 is AM-TMP

consultants-6 is agent

consultants-6 participating-in IDSAEIN-12

consultants-6 typeOf infectious_diseases_consultants

IDSAEIN-12 is participating-in-what?

consultants-6 participating-more-detail currently-8

The modified T2R triples:

{ IDSAEIN, is, Loc . }

{ 300 consultants, is, Agent . }

{ 300 consultants, participate in, IDSAEIN . }

{ infectious diseases consultants, typeOf, consultants . }

{ IDSAEIN, is, Participating-in-what? . }

{ participate, when, currently . }

{ 300 consultants, participate, IDSAEIN . }

{ 300 consultants, typeOf, consultants . }

Modification of triples with the property “more-detail” for nouns: In some cases, the “more-detail” property provides some information about adjectives. In such situations, new triples as shown in the example below are created to make the results simpler and more meaningful.

The analyzed sentence: *"it is too late."*

T2R triples:

late-4 subject it-1

late-4 more-detail too-3

The new Triples:

{ it, is, too late . }

{ too late, typeOf, late}

As we can see, the object of the triple with the "more-detail" property (i.e. "too") is appended to any occurrences of the subject of triple with the property "more-detail" (i.e. "late") in other triples. A new triple with the property "typeOf" (i.e. "{ too late, typeOf, late} ") is created to keep the output simple and self-explanatory.

Modification of triples with the property "verb + direct-object": In some cases, a phrase "direct-object" is a part of the properties of some T2R triples. In such situations, the other part of such a property is a verb. The object of the triple that contains the "direct-object" part is the actual object of the corresponding verb, and the subject of this triple is the actual subject of that verb. As seen in the previous cases, those subject and object nodes are only single words. If "typeOf" or any other conjunction of nouns related to those subjects and objects (e.g. "of") exist, their complete nouns phrases are identified and those single words are replaced with the complete noun phrases. This replacement makes the triples more meaningful. The "typeOf" triples are still kept to maintain the simplicity of the whole set of triples. The example of this case is as follows:

The analyzed sentence: *"The current data collection covers 11 zoonotic agents."* [24]

T2R triples:

collection-4 **covers-direct-object** agents-8

collection-4 typeOf current_data_collection

collection-4 is instrument

agents-8 is thing-covered

agents-8 typeOf zoonotic_agents

agents-8 number 11-6

The modified triples:

{ current data collection, cover, 11 agents . }

{ current data collection, typeOf, collection . }

{ current data collection, typeOf, data collection . }

{ data collection, typeOf, collection . }

{ current data collection, is, Instrument . }

{ 11 agents, is, Thing-covered . }

{ zoonotic agents, typeOf, agents . }

{ coverer, cover, 11 agents . }

{ 11 agents, typeOf, agents . }

Generation of triples from sentences with negative meaning: Time to time, sentences have negative meaning. There are a few "negative patterns", it means that the negation is applied to a different part of sentence. For each such negation pattern, different modifications of triples are required. A number of possible patterns are shown in the following examples.

Example 1:

The analyzed sentence: *"the perspective does not permit prediction."*

T2R triples:

perspective-2 is allower

not-4 is AM-NEG

perspective-2 not-permit-direct-object prediction-6

prediction-6 is action-allowed

In order to cover the negative meaning in the above case, the negative word (i.e. "not") is placed before the corresponding verb as shown in the following T2R triples:

The modified triples:

{ perspective, is, Allower . }

{ not, is, Am-neg . }

{ perspective, **not permit**, prediction . }

{ prediction, is, Action-allowed . }

{ perspective, **not permit**, prediction . }

Example 2:

The analyzed sentence: *"The end of the epidemic is not yet in sight."* [32]

T2R triples:

end-2 not-is-in sight-10

end-2 of epidemic-5

end-2 not-is-more-detail yet-8

The modified triples:

{ end, **is not**, sight . }

{ end, **is not**, yet . }

{ end of epidemic, **is not**, sight . }

{ end of epidemic, **is not**, yet . }

{ end of epidemic, typeOf, epidemic .}

In the above case, the sentence contains a negative "to be" (i.e. "is") verb. Therefore, the negative word "not" is placed after the verb "to be".

Example 3:

The analyzed sentence: *"No probable case of SARS was diagnosed among these patients."* [35]

T2R triples:

patients-10 is illness

case-3 is doctor

case-3 typeOf probable_case

SARS-5 is ORG

case-3 of SARS-5

case-3 diagnosed-among patients-10

In this example, the sentence contains the word "no" as a determiner. The T2R tool is not able to handle these kinds of negative sentences correctly in most cases. Therefore, to bring the negative meaning to the triples, the "no" word is searched using the Senna's POS and the location of the word in any given sentence. Then, if the sentence has the word "no", it is placed

before any occurrence of the corresponding word in the generated RDF triples. In addition, other triples with the property "typeOf" are created, as it is seen in the modified triples below.

The modified triples:

{ patients, is, Illness . }

{ probable case, is, Doctor . }

{ probable case, typeOf, case . }

{ SARS, is, Org . }

{ no probable case, diagnosed among, patients . }

{ **no probable case**, **typeOf, case** . }

{ **no probable case**, **typeOf, probable case** .}

Modification of triples with the property "subject": In some cases, the T2R triples contain triples with "subject" property as seen in the example below:

The analyzed sentence: *"Obama is a graduate of Columbia University and Harvard Law School."*

T2R triples:

Obama-1 is PER
graduate-4 of ColumbiaUniversityandHarvardLawSchool-6
ColumbiaUniversityandHarvardLawSchool-6 is LOC
graduate-4 subject Obama-1

A portion of the modified triples:

{ Obama, is, Per . }

{ ColumbiaUniversityandHarvardLawSchool, is, Loc . }

{ Obama, is, graduate of ColumbiaUniversityandHarvardLawSchool . }

As seen in the above example, if any verb of a sentence that has a relationship with the subject of the triple with the property "subject" is found by utilizing Senna's POS and semantic role labeling, that verb (i.e. "is") is considered as a property of the new triple.

Generation of triples with the property "typeOf" from triples with the property "of": If a triple with the property "of" exist in T2R triples then a new triple with the property "typeOf" and the following nodes is created:

Subject: is the subject of the triple with the property "of" together with a word "of" and the object of this triple

Property: "typeOf"

Object: is the object of the triple with property "of"

The following example demonstrates the application of this case:

The analyzed sentence: *"Another visible accomplishment is the elimination of hydatidosis in the endemic countries and regions of the southern cone."* [22]

A portion of the T2R triples:

elimination-6 of hydatidosis-8
hydatidosis-8 in regions-14
countries-12 of cone-18

A portion of the modified triples:

{ visible accomplishment, is, elimination of hydatidosis . }

{ **elimination of hydatidosis**, typeOf, hydatidosis . }

3.8 Utilizing Stanford Co-reference Chain

In this study, the Stanford Deterministic Co-reference Resolution system [36] is used to find a reference of any pronoun that is present in the generated RDF triples from a given sentence. The Stanford CoreNLP library is built upon Java codebases. Therefore, the Stanford NLP (edu.stanford.nlp.*) java libraries are used in order to find the references of all pronouns.

Dependency Co-reference resolves how a word implies another word in sentences, such as pronouns that refer to individuals. Dependency Co-reference tries to provide a chain, which demonstrates how phrases refer to each other and which phrase is representative, that creates clusters [37].

The following example shows the representative mentions and their references for a given sentence.

The analyzed sentence: *"London had an official population of 8,174,100 , making **it** the most populous municipality in the European Union, and accounting for 12% of the UK population."*

ClusterID: 1

Representative Mention: London,

Position: HeadWord: 1 [1,2],

Sentence number: 1

Mentions:

it, - **Position** 10 [10,11] - **Sentence number**: 1

As shown in the above example, the *ClusterID* represents all related phrases. The *Representative Mention* is "London" in the above sentence. In this case, "London" is found in sentence 1. The *Position* indicates the location of the word in the sentence with the *startIndex* and *endIndex*. The *endIndex* is always one word prior the start. In this example, it is just the one word, "London", found as the first token in the sentence. The *HeadWord* is an offset of the most important word of the phrase [37]. For instance, in the phrase "Another visible accomplishment", "another" and "visible" both explain "accomplishment" which is the core of the phrase.

It should not be considered any representative mention that only have one mention which referring to itself, otherwise in this case the representative mention and the mention are same [37].

The utilization of output of the Stanford co-reference chain is explained with an example presented below.

The analyzed sentence: *"London is a big city in the southeast of England, on the River Thames. It is the capital of England and the United Kingdom."*

RDF Triples created with Boxer tool:

{ neuter, is, capital of England . }

{ **neuter, Coreference, London** . }

{ capital of England, typeOf, England . }

{ capital of England, typeOf, capital . }

{ England, type, loc . }

{ london, type, loc . }

{ UnitedKingdom, type, org . }

RDF Triples created with T2R tool:

{ UnitedKingdom, is, Misc . }

{ England, is, Loc . }

{ it, is, capital of UnitedKingdom . }

{ it, Coreference, London . }

{ it, is, capital of England . }

As seen in the above example, we created a RDF triple (by utilizing The Stanford CoreNLP library) with the property "Corefrence". However, if the given input is a paragraph, there might be same pronouns that refer to different references. One way to solve this issue is to replace all the nodes with words "neuter", "person" or "thing" that represents pronouns found in C&C-Boxer RDF triples with the object of the generated RDF triple with the property "Corefrence". In addition, all the pronouns in T2R triples can be replaced as well with the object as shown in the following triples.

The triples generated by Boxer tool:

{ London, is, capital of England . }

{ capital of England, typeOf, England . }

{ capital of England, typeOf, capital . }

{ England, type, loc . }

{ london, type, loc . }

{ UnitedKingdom, type, org . }

The triples generated by T2R tool:

{ UnitedKingdom, is, Misc . }

{ England, is, Loc . }

{ London, is, capital of UnitedKingdom . }

{ London, is, capital of England . }

In some cases, the representative mention can be a long noun phrase, as seen in the following example. In such a situation, if more than one mention exist for the representative mention, those

mentions will be investigated to see if a short mention could be found and used to replace the corresponding pronouns in the triples.

The analyzed sentence: *"Conservation medicine, the medical practice that seeks to promote ecological health and well being of a defined habitat, functions at the intersection of animal, human and ecosystem health. It differs from classical public-health epidemiology and medicine in that it aims to protect and improve animal health and related ecosystems, in addition to human health."* [38]

clusterID 1
" **Representative Mention**: the medical practice that seeks to promote ecological health and well being of a defined habitat -Position 6 [4,20] - Sentence number: 1

" **Mentions**:
Conservation medicine, the medical practice that seeks to promote ecological health and well being of a defined habitat - Position 2 [1,21] - Sentence number: 1

" **Mentions**:
Conservation medicine - Position 2 [1,3] - Sentence number: 1

" **Mentions**:
It - Position 1 [1,2] - Sentence number: 2

" **Mentions**:
it - Position 11 [11,12] - Sentence number: 2

As seen in the above example, the representative mention of the pronoun "it" is "*the medical practice that seeks to promote ecological health and well being of a defined habitat*" which is a long-term phrase. It is not desirable to place such a long phrase instead of the pronouns existing in the triples. However, among the mentions, the actual and short reference of the pronoun "it" can be found – it is "Conservation medicine" noun phrase.

3.9 Visualization

In order to store and provide access to the generated RDF triples, the Virtuoso Universal Server [39] is utilized in this study. "*Virtuoso comes with a web-based application called Virtuoso Conductor that provides an interface for the database management functionality"* [40] generally performed by data manger. We utilize virtuoso.jena libraries to work with this Virtuoso Conductor using Java programs.

In order to visualize the RDF structure in the form of RDF graph, the Gephi tool [41] can be utilized.

Gephi is an open-source software package written in java for network analysis and visualization. There is a Gephi plug-in called virtuosoimporter [42] which allows importing of RDF data from a Virtuoso server. It uses the import spigot functionality. In this study, generated RDF triples stored in the Virtuoso Conductor can be visualized using this plugin.

4. Evaluation

4.1 Evaluation Criteria

The evaluation of "goodness" of generated RDF triples is not a straightforward process. So far, there is no clear approach of doing it. In most of the cases, humans evaluate triples, i.e., they subjectively assess quality of triples based on their comparison with the text used for generation of the triples.

In this study a novel approach is proposed. It does not need any involvement of humans. It is based on a number of criteria that are imposed on RDF triples. There are four criteria we define: simplicity, coverage, clearness, and confidence. A set of RDF triples is evaluated via determining levels of satisfaction of those criteria by these triple.

Simplicity

According to its formal definition simplicity is:

"… the state or quality of being simple. Something which is easy to understand or explain is simple." [43].

Based on this definition, we have determined a phrase "simple RDF triple" in the following way: if RDF's nodes, i.e., subject, property, object, are simple words then the RDF triple meets the simplicity criterion.

Of course, some of generated triples satisfy this criterion, but there are triples which are complex, i.e., they have complex nodes. A complex node means a node that is a noun phrase containing different parts of speech, e.g. a noun phrase with different kind of adjectives, adverbs before adjectives, articles and more than one noun. The general noun phrase pattern is shown below[10]:

Noun Phrase pattern: $(M_1)_3(M_2)_\infty(M_3)_2$ N

where: M_1 identifies modifiers – there could be a max of three modifiers, M_2 adjectives –

[10] Dr. Amookhteh's lecture (in Farsi)

unlimited number, M_3 nouns – max of two, and N is a core noun.

In general, nodes of a triple can have different kinds of modifiers (M_1). Their list is below:

- an article: “a”, “an”, “the”;
- a demonstrative adjective: “this”, ”these”, ”such”, etc.;
- quantifiers
  - such as a different kind of intensifiers: ”too”, “so”, ”very”;
  - “some”, “any”, “many”, ”several”, etc.;
  - ordinal and cardinal numbers: “two”, “the second”, etc.

Since C&C-Boxer is able to eliminate articles and demonstrative adjectives, there is no need to perform any simplification tasks. However, in the case of quantifiers we need to simplify the nodes as explained later.

Different kinds of adjectives (M_2) can play the role of another element of the noun phrase. The element before the core noun could be a noun phrase (M_3) that cannot contain more than two nouns.

The following example shows a set of triples generated from a given sentence with a complex noun phrase. The sentence is:

*“A way would complement **other public health surveillance efforts.**”* [31]

and it contains the noun phrase:

Other (M_1) public (M_2) health (M_2) surveillance (M_3) efforts (N)

RDF triples generated from the sentence are:

{way, complement, public health surveillance efforts . }
{ public health surveillance efforts, typeOf, efforts . }
{ public health surveillance efforts, typeOf, public efforts . }
{ public health surveillance efforts, typeOf, public surveillance efforts . }
{ public health surveillance efforts, typeOf, health surveillance efforts . }
{ health surveillance efforts, typeOf, surveillance efforts . }
{surveillance efforts, typeOf, efforts . }

{ public efforts, typeOf, efforts . }

As seen in the above triples, a number of triples contain the property "typeOf". These triples refer to the core noun of the noun phrase "public health surveillance efforts".

In order to evaluate simplicity of an RDF triple, we define a "simplicity ratio". The ratio is determined only based on triples without the property "typeOf". A score of each triple is calculated using the scores assigned to each node, i.e., subject and object. The maximum score for a single triple is one, and each node can have a score of zero or point five. The nodes' scores are determined in the following way:

If both subject and object of a given triple are simple, i.e., they contain only one word each – the score for each node is 0.5. If a node (subject, object) is not simple but there are other triples related to this node that have "typeOf" property and their subjects, objects "explain" this node, then the score of 0.5 is also given to this node. Otherwise the score is zero.

Overall, the simplicity ratio is equal to the average of triples' scores. Its value is calculated based on the following formula:

$$\textbf{Simplicity ratio} = \frac{\sum \textbf{triples' scores}}{\textbf{Total number of generated triples (without "typeOf" property)}}$$

Coverage

In the case of coverage measure, we define two types of coverage: essence coverage, and word coverage. The details about each of them are given below.

Essence Coverage

The essence coverage criterion is evaluated to determine to what extend generated RDF triples cover the essence of a sentence. In practice, it means that main verbs of a given sentence and their dependencies are included in RDF triples. The essence of the sentences is covered when:

- verbs of a given sentence are properties of triples, and those triples' subjects and objects contain the dependencies or the arguments of the verbs as defined in the PropBank [18];

- a given sentence is in the passive voice format or the tools used to generate triples fail to recognize the arguments of the verbs; and the verbs of the sentence exist in the subject of the triple.

In the first case, the scores of 0.5 are assigned to subjects and objects. If subjects and objects are correctly identified according to verb's arguments, the scores of 0.5 are given. Otherwise, it is zero. For example, if a triple only recognizes the correct subject, the triple's score is 0.5.

In the second case, there is no subject in the triple because of the passive voice; a score related to the subject recognition is reduced to zero. For determining the score related to the recognition of the object, the triples are evaluated – checked against the arguments from PropBank.

It should be noted that in order to identify all main verbs of a given sentence, we use Senna and Boxer that are able to recognize Parts of Speech (POS). This leads to identification of roles of elements in given sentences.

The essence coverage ratio is equal to the average of triples' scores. It is calculated using the following equation:

$$\textbf{Essence coverage ratio} = \frac{\sum \textbf{the score of the subject and the object for each triple containing verbs}}{\textbf{Total Number of verbs}}$$

Word Coverage

This criterion is evaluated to determine if words of a given sentence are observed at least once in the generated RDF triples. The word coverage ratio is calculated as follows:

$$\textbf{Words Coverage ratio} = \frac{\textbf{Number of words found in the generated triples}}{\textbf{Total number of words of a given sentence}}$$

In a given sentence, any individual word is counted once while the words that are repeated are not counted. It should be emphasized that some words such as articles (modifier), conjunctions and proposition are not considered while their eliminations do not decrease the meaning of sentences.

Clearness

This criterion is evaluated to check whether the generated RDF triples contain phrases that are understandable. In the other words, it evaluates whether the triples are well explained. In this study, we use two methods to sure that generated triples are well explained.

One of these methods is using Stanford Co-reference Chain Tool [11]. It determines a meaning of any pronoun existing in the triple, and replaces such a pronoun with its corresponding reference.

Another method eliminates acronyms. For each acronym existing in a given triple, it creates a triple with the property "acronym for". This triple explains the meaning of the acronym.

The ratio that represents clarity is very simple:

$$\textbf{Clarity ratio} = \frac{\textbf{Number of resolved pronouns or acronyms}}{\textbf{Total Numbers ofpronouns or acronyms}}$$

Confidence level

This criterion is used to evaluate whether two sets of generated RDF triples – one set produced by C&C-Boxer and anther by T2R – are similar. The similarity of triples is determined by investigation of triples, from both sets, that contain main verbs of a given sentence as their properties. In other words, the subjects and objects of triples that have verbs as their properties are compared. The motivation behind this criterion is the fact that verbs and their arguments reflect the essence of a sentence. Therefore, if the triples generated by different tools have similar/same nodes then this demonstrates that different tools can find essential parts of the sentence. This leads to increased confidence in the generated RDF triples.

Defining similarity score: The approach adopted in this study to evaluate the similarity of the triples' nodes generated by both tools is described as follows:

For triples, from both sets, that contains a main verb of a given sentence as the property: their subjects and objects are compared. If these nodes are different, the other triples are investigated to check whether another relationship or similarity can be found between them. If similar nodes are identified, the score of 0.5 is given to each node. A total score for similar triples

is 1.0.

For triples with a verb is found as the subject node instead of the property: this situation requires checking similarity only between objects of the triples. In this case the similarity score for subjects is 0.0 since they are not considered.

It is possible that the main verb(s) might be missing in the triples generated by one of the tools. Consequently, no triple exists to be compared. In such situations, the score of 0.0 will be given as the confidence level for triples with the verbs generated by the other tool.

The confidence level ratio is calculated using the following equation:

$$\textbf{Confidence Level Ratio} = \frac{\sum \textbf{the similarity score between the triples including the main verb}}{\textbf{Total number of main verbs}}$$

In the following subsections, we include results of tests we have performed on a number of sentences. These sentences have been randomly selected from forty-nine abstract of research papers in the domain of agriculture.

The performed studies are divided into three groups. The sentences from each group are tested for different criteria as shown in the Table 4.1. It should be emphasized that for the criteria simplicity, coverage, and confidence level (Case Study A), the tested sentences are the same. For the case studies B and C the sentences are different.

Table 4.1 Details of Conducted Studies

Case Study	Criteria	No of Sentences
A	Simplicity, Essence Coverage, Word Coverage, Confidence Level	10
B	Co-reference Chain check	5
C	Acronym Check	5

4.2 Evaluation Case Studies

4.2.1 Case Study A

Sentence_1:

"Nevertheless, privatization of animal and human health services has had a negative effect on human resources and infrastructure by weakening essential epidemiological functions in some countries." [22]

Boxer Results:

The RDF triples generated by Boxer used on the Sentence_1 are shown in Table 4.2.

Table 4.2 Sentence_1: Boxer Output

RDF triples generated using Boxer tool	Simplicity Score
{ privatization of animal, have, negative effect . }	1
{ human health services, have, negative effect . }	1
{ negative effect, on, human resources . }	1
{ negative effect, on, infrastructure . }	1
{ have, how, by weaken . }	1
{ weaken, patient, essential epidemiological functions . }	1
{ essential epidemiological functions, in, countries . }	1
{ privatization, of, animal services.}	1
{ privatization, of, human services . }	1
{ essential epidemiological functions, typeOf, functions . }	-
{ privatization of animal, typeOf, animal . }	-
{ negative effect, typeOf, effect . }	-
{ privatization of animal, typeOf, privatization . }	-
{ human health services, typeOf, human . }	-
{ human health services, typeOf, human services . }	-
{ human resources, typeOf, resources . }	-
{ essential epidemiological, typeOf, epidemiological . }	-
{ human services, typeOf, services . }	-

Simplicity Ratio=9/9=100%

As seen in Table 4.2, the simplicity score of each triple except for triples with "typeOf" property equals to 1.0. It is due the fact that, for each triple, not only the subject but also the object of the triple are either simple or have corresponding triples with "typeOf" property that

shows the core of the noun phrase.

Essence Coverage Ratio = (0.5+ 0.5)/1=100%

As seen in the following RDF triples, they cover the essence of the sentence. The triples contain the main verb of the sentence “have” and their arguments. Therefore, the Boxer essence coverage equals 100%:

Triples with the verb “have”:

{ privatization of animal, **have**, negative effect . }
{ human health services, **have**, negative effect . }
{ privatization, of, human services . }
{ human health services, typeOf, human services . }
{ human services, typeOf, services . }

Words Coverage Ratio=16/16=100%

The words of the sentence are listed as follows: “1-privatization, 2-animal, 3-human, 4-health, 5- services, 6- had (have),7- negative, 8- effect, 9- human , 10- resources , 11-infrastructure, 12-weaken, 13-essential, 14- epidemiological , 15-functions , 16-countries.”

In the generated RDF triples, it is observed that all the above words appeared at least once, therefore, the words coverage of this case study equals 100%.

T2R Results:

The analysis of the Sentence_1 by T2R tool leads to generation of a set of RDF triples as shown in Table 4.3.

Table 4.3 Sentence_1: T2R Output

RDF triples generated using T2R tool	Simplicity Score
{ privatization, have, negative effect . }	1
{ human health services, have, negative effect . }	1
{ privatization of animal, have, negative effect . }	1
{ negative effect, on, infrastructure . }	1
{ anonymous, weaken, essential epidemiological functions . }	1
{ weaken, in, countries . }	1
{ negative effect, on, human resources . }	1
{ privatization of animal human health services, have, negative effect . }	1
{ negative effect, typeOf, effect . }	-
{ resources human, typeOf, resources . }	-
{ privatization of animal, typeOf, animal . }	-
{ resources human, typeOf, resources . }	-
{ human health services, typeOf, services . }	-
{ essential epidemiological functions, typeOf, functions . }	-
{ human resources, typeOf, resources . }	-
{ animal human health services, typeOf, services . }	-
{ privatization of animal human health services, typeOf, animal human health services . }	-

Simplicity Ratio=8/8=100%

The simplicity ratio of the RDF triples generated by T2R is the same as in the case of Boxer. That is, for each triple, not only the subject but also the object of the triple are either simple or have corresponding triples with "typeOf" property that shows the core of the noun phrase.

Essence Coverage Ratio = (0.5+ 0.5)/1=100%

As seen in the following selected T2R RDF triples, these RDF triples cover the essence of the sentence, i.e., they contain the main verb of the sentence. Therefore, the T2R essence coverage equals 100%.

Triples with the verb "have":

{ privatization of animal, **have**, negative effect . }
{ human health services, **have**, negative effect . }
{ privatization of animal human health services, have, negative effect . }

Words Coverage Ratio=16/16=100%

The word coverage ratio equals 100% since all the listed words appeared at least once in the

generated RDF triples.

Confidence level:

Table 4.4 Sentence_1: RDF Triples with the main verb

Tools	Main Verb (property)	Subject	Property	Object
Boxer	have	privatization of animal	have	negative effect
		human health services	have	negative effect
T2R	have	privatization of animal	have	negative effect
		human health services	have	negative effect

As seen in Table 4.4, the triples with "have" property have the same subjects and objects in the RDF triples generated by both tools. That is, the score of 0.5 is given to the subjects and the score of 0.5 to the objects. The confidence score equals 1.0, which indicates that the generated RDF triples have 100% confidence level:

$$\textbf{Confidence Level Ratio} = \frac{\sum \textbf{the similarity score between the triples including the main verb}}{\textbf{Total number of main verbs}}$$

$$= \textbf{(0.5+0.5) /1=100\%}$$

The values of measures obtained for the Sentence_1 are summarized in Table 4.5.

Table 4.5 Summary of the results for the Sentence_1

Simplicity		Essence Coverage		Word Coverage		Confidence Level
Boxer	T2R	Boxer	T2R	Boxer	T2R	
100%	100%	100%	100%	100%	100%	100%

Sentence_2:

"Health care reform in the USA provides an opportunity to address critical needs, such as improved surveillance and diagnosis, to ensure timely detection of and rapid response to newly emerging infectious diseases." [23]

Boxer Results:

The RDF triples generated by Boxer used on the Sentence_2 are shown in Table 4.6.

Table 4.6 Sentence_2: Boxer Output

RDF triples generated using Boxer tool	Simplicity Score
{ critical needs, ensure, timely detection . }	1
{ care reform, in, Usa . }	1
{ critical needs, as, improved diagnosis surveillance . }	1
{ opportunity, address, critical needs . }	1
{ care reform, provides, opportunity . }	1
{ timely detection of rapid response, to, emerge infectious diseases newly . }	1
{ diagnosis, ensure, timely detection . }	1
{ diagnosis, ensure, timely detection of rapid response . }	1
{ usa, type, org . }	1
{ Health, type, nam . }	1
{ reform, in, USA . }	1
{ critical needs, as, improved surveillance diagnosis . }	1
{ rapid response, to, newly emerge infectious diseases . }	1
{ reform, provides, opportunity . }	1
{ timely detection, of, rapid response . }	1
{ Health care reform, typeOf, reform . }	-
{ care reform, typeOf, reform . }	-
{ Health care reform, typeOf, care reform . }	-
{ critical needs, typeOf, needs . }	-
{ improved diagnosis, typeOf, diagnosis . }	-
{ improved diagnosis surveillance, typeOf, surveillance . }	-
{ timely detection, typeOf, detection . }	-
{ timely detection of rapid response, typeOf, rapid response . }	-
{ timely detection of rapid response, typeOf, timely detection . }	-
{ timely detection, typeOf, detection . }	-
{ rapid response, typeOf, response . }	-
{ emerge infectious diseases, typeOf, diseases . }	-
{ newly emerge infectious diseases, typeOf, newly . }	-
{ care, typeOf, care reform . }	-
{ Health, typeOf, Health reform . }	-
{ reform, typeOf, care reform . }	-
{ reform, typeOf, Health reform . }	-

Simplicity Ratio = 15/15=100%

As seen in Table 4.6, the simplicity score of each triple except for triples with "typeOf" property equals to 1.0. It is due the fact that, for each triple, not only the subject but also the object of the triple are either simple or have corresponding triples with "typeOf" property that shows the core of the noun phrase.

Essence Coverage Ratio = (0.5+ 0.5) + (0.5+0.5)/2=100%

As seen in the following selected RDF triples, they cover the essence of the sentence. The triples contain the main verbs of the sentence including "provide", "ensure", and the arguments of the verbs. Therefore, the Boxer essence coverage equals 100%:

Triples with the verb "provide":

{ care reform, **provides**, opportunity . }
{ Health care reform, typeOf, care reform . }
{ opportunity, address, critical needs . }

Triples with the verb "ensure":

{ critical needs, **ensure**, timely detection . }
{ diagnosis, **ensure**, timely detection . }
{ diagnosis, **ensure**, timely detection of rapid response . }

Words Coverage Ratio=21/21=100%

The words of the sentence are listed as follows: "1-Health, 2-care, 3- reform, 4-USA, 5-provides, 6- opportunity, 7- address, 8- critical, 9- needs, 10-improved, 11- surveillance, 12-diagnosis, 13- ensure, 14-timely, 15-detection, 16- rapid, 17-response, 18-newly, 19-emerging, 20- infectious, 21- diseases."

In the generated RDF triples, it is observed that all the above words appeared at least once, therefore, the words coverage of this case study equals 100%.

T2R Results:

The analysis of the Sentence_2 by T2R tool leads to generation of a set of RDF triples as shown in Table 4.7.

Table 4.7 Sentence_2: T2R Output

RDF triples generated using T2R tool	Simplicity Score
{ Health care reform, provide, critical needs . }	1
{ critical needs, improved, surveillance . }	1
{ Health care reform, in, USA . }	1
{ Health care reform, provide, detection . }	1
{ Health care reform, provide, surveillance . }	1
{ Health care reform, provide, opportunity . }	1
{ USA, is, Loc . }	1
{ opportunity, address, critical needs . }	1
{ opportunity, address, surveillance . }	1
{ critical needs, improved, surveillance . }	1
{ critical needs, improved, diagnosis . }	1
{ provide, when, newly . }	1
{ Health care reform, provide, infectious diseases . }	1
{ Health care reform, in, USA . }	1
{ address, is, Critical . }	1
{ opportunity, is, critical needs . }	1
{ critical needs, improved, rapid response diagnosis . }	1
{ rapid response diagnosis, typeOf, response . }	-
{ infectious diseases, typeOf, diseases . }	-
{ diseases-response, typeOf, response . }	-
{ diagnosis rapid response, typeOf, response . }	-
{ diseases infectious, typeOf, diseases . }	-
{ critical needs, typeOf, needs . }	-
{ Health care reform, typeOf, reform . }	-
{ critical needs, typeOf, needs . }	-

Simplicity Ratio =17/17=100%

The simplicity ratio of the RDF triples generated by T2R is the same as in the case of Boxer. That is, for each triple, not only the subject but also the object of the triple are either simple or have corresponding triples with "typeOf" property that shows the core of the noun phrase.

Essence Coverage Ratio = (0.5+ 0.5) + (0+0)/2=50%

As seen in the following selected T2R RDF triples, these RDF triples cover the verb "provide" but no triple exists to contain the verb "ensure". That is, the T2R essence coverage of this case study equals 50%.

Triples with the verb "provide":

{Health care reform, **provide**, opportunity.}

Words Coverage Ratio: 18/21= 85%

The generated RDF triples do not contain the words including "ensure", "timely" and "emerge", therefore, the word coverage ratio equal 85%.

Confidence level:

Table 4.8 Sentence_2: RDF Triples with the main verb

Tools	**main verb (property)**	**subject**	**Property**	**Object**
Boxer	provide	care reform	provide	opportunity
		Health care reform	typeOf	care reform
T2R	provide	Health care reform	provide	opportunity
Boxer	ensure	critical needs	ensure	timely detection
		Diagnosis	ensure	timely detection .
		Diagnosis	ensure	timely detection of rapid response
T2R	ensure	"NO similar nodes"	"NO similar nodes	"NO similar nodes"

As seen in Table 4.8, there is only similarity between subjects and objects of triples containing the verb "provide" in the RDF triples generated by both tools. RDF triples provided by T2R do not include any triples with the verb "ensure" as a node. Therefore, the confidence score equals 0.5, which indicates that the generated RDF triples have 50% confidence level:

$$\textbf{Confidence Level Ratio} = \frac{\textbf{the similarity score between the triples including the main verb}}{\textbf{Total number of main verbs}}$$

$$= \textbf{(0.5+0.5) + (0+0)/ 2 =50\%}$$

The values of measures obtained for the Sentence_2 are summarized in Table 4.9.

Table 4.9 Summary of the results for the Sentence_2

Simplicity		**Essence coverage**		**Word coverage**		**Confidence level**
Boxer	**T2R**	**Boxer**	**T2R**	**Boxer**	**T2R**	
100%	**100%**	**100%**	**50%**	**100%**	**85%**	**50%**

Sentence_3:

"Detection and surveillance of emerging zoonoses have greatly benefited from technical progress in diagnostics." [44]

Boxer Results:

The RDF triples generated by Boxer used on the Sentence_3 are shown in Table 4.10.

Table 4.10 Sentence_3: Boxer Output

RDF triples generated using Boxer tool	Simplicity Score
{ benefit, how, greatly . }	1
{ benefit, agent, Detection . }	1
{ technical progress, in, diagnostics . }	1
{ Detection, benefit from, technical progress . }	1
{ benefit, agent, surveillance . }	1
{ benefit, agent, surveillance of emerge zoonoses . }	1
{ technical, in diagnostics, progress . }	1
{ technical, in diagnostics, progress . }	1
{ benefit , agent, Detection . }	1
{ technical progress, in, diagnostics . }	1
{ surveillance, benefit from, technical progress . }	1
{ surveillance, of, emerge zoonoses . }	0.5
{ benefit greatly, typeOf, benefit . }	-
{ technical progress, typeOf, progress . }	-
{ surveillance of emerge zoonoses, typeOf, emerge zoonoses . }	-
{ surveillance of emerge zoonoses, typeOf, surveillance . }	-
{ technical progress, typeOf, progress . }	-

Simplicity Ratio = 11.5/12 =95 %

As calculated above, the simplicity ratio of the generated RDF triples utilizing Boxer output on Sentence_ 3 equals 95% since the object of one of the triple neither is a simple word nor has a related triple with the property "typeOf".

Essence Coverage Ratio = (0.5+ 0.5)/ 1=100%

As seen in the following selected RDF triples, they cover the essence of the sentence. The triples contain the main verb of the sentence "benefit" and its arguments. Therefore, the Boxer essence coverage equals 100%.

Triples with the verb "benefit":

{ Detection, **benefit from**, technical progress . }

{ surveillance, **benefit from**, technical progress . }
{ surveillance, of, emerge zoonoses . }

Words Coverage Ratio =9/9=100%

The words of the sentence are listed as follows: "1-Detection, 2- surveillance, 3- emerging, 4-zoonoses, 5-greatly, 6-benefited, 7- technical, 8- progress, 9-diagnostics."

In the generated RDF triples, it is observed that all the above words of the sentence appeared at least once. As a result, the word coverage ratio of this case study equals 100%.

T2R Results:

The analysis of the Sentence_3 by T2R tool leads to generation of a set of RDF triples as shown in Table 4.11.

Table 4.11 Sentence_3: T2R Output

RDF triples generated using T2R tool	Simplicity Score
{ benefit, how, greatly . }	1
{ Detection, benefit from, technical progress . }	1
{ emerging zoonoses, benefit from, technical progress . }	1
{ Detection, benefit in, diagnostics . }	1
{ surveillance, of, emerging zoonoses . }	1
{ emerging zoonoses, benefit in, diagnostics . }	1
{ emerging zoonoses, benefit from, technical progress . }	1
{ Detection, of, emerging zoonoses . }	1
{ emerging zoonoses, benefit in, diagnostics . }	1
{ Detection surveillance zoonoses, typeOf, zoonoses . }	-
{ technical progress, typeOf, progress . }	-
{ technical progress, typeOf, progress . }	-
{ emerging zoonoses, typeOf, zoonoses . }	-
{ technical progress, typeOf, progress . }	-
{ surveillance zoonoses, typeOf, zoonoses . }	-
{ emerging zoonoses, typeOf, zoonoses . }	-

Simplicity Ratio = 9/9=100 %

As calculated above, the simplicity ratio of the produced RDF triples utilizing T2R output on Sentence_3 equals 100%. It is due the fact that, for each triple, not only the subject but also the object of the triple are either simple or have corresponding triples with "typeOf" property that shows the core of the noun phrase.

Essence Coverage Ratio = (0.5+ 0.5) /1=100%

As seen in the following selected RDF triples, the essence of the sentence (main verb and its dependencies) appeared in the generated T2R RDF triples. Therefore, the T2R essence coverage equals 100%.

Triples with the verb "benefit":

{ Detection, **benefit from**, technical progress . }
{ Detection, of, emerging zoonoses . }
{ emerging zoonoses, **benefit from**, technical progress . }
{ surveillance, of, emerging zoonoses . }

Words Coverage Ratio =9/9=100%

The word coverage ratio equals 100% since all the listed words appeared at least once in the generated RDF triples.

Confidence level:

Table 4.12 Sentence_3: RDF Triples with the main verb

Tools	**main verb(property)**	**subject**	**Property**	**Object**
Boxer	Benefit	Detection	benefit from	technical progress
		surveillance	benefit from	technical progress
		surveillance	of	emerge zoonoses
T2R	benefit	Detection	benefit from	technical progress
		emerging zoonoses,	benefit from	technical progress
		surveillance	of	emerging zoonoses

As seen in Table 4.12, triples with "benefit" property have similar or the same subjects and objects in the RDF triples generated by both tools. That is, the score of 0.5 is given to the subjects and the score of 0.5 to the objects. The confidence score equals 1.0, which indicates that the generated RDF triples have 100% confidence level:

$$\textbf{Confidence Level Ratio} = \frac{\sum \textbf{ the similarity score between the triples including the main verb}}{\textbf{Total number of main verbs}}$$

$$= \textbf{(0.5+0.5)/1 =100\%}$$

The values of measures obtained for the Sentence_3 are summarized in Table 4.13.

Table 4.13 Summary of the results for the Sentence_3

Implicity		Essence coverage		Word coverage		Confidence level
Boxer	**T2R**	**Boxer**	**T2R**	**Boxer**	**T2R**	
95%	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

Sentence_4:

"*Overcoming these challenges and limitations will require a concerted effort from a variety of sources, including an ongoing partnership between infectious disease clinicians and public health professionals."* [45]

Boxer Results:

The RDF triples generated by Boxer used on the Sentence_4 are shown in Table 4.14.

Table 4.14 Sentence_4: Boxer Output

RDF triples generated using Boxer tool	Simplicity Score
{ concerted effort, from, variety . }	1
{ concerted effort, from, variety of sources . }	1
{ variety, of, sources, . }	1
{ overcome, require, concerted effort . }	1
{ overcome, patient, challenges . }	1
{ require, patient, concerted effort . }	1
{ variety of sources, include, ongoing partnership . }	1
{ ongoing partnership, between, infectious disease clinicians . }	1
{ ongoing partnership, between, infectious clinicians public health professionals . }	1
{ limitations, require, concerted effort . }	1
{ ongoing partnership, between, infectious clinicians public professionals . }	0.5
{ sources,, include, ongoing partnership . }	1
{ thing, overcome, challenges . }	1
{ health, typeOf, health infectious clinicians public professionals . }	-
{ infectious clinicians public health professionals, typeOf, professionals . }	-
{ infectious disease clinicians public health professionals, typeOf, professionals . }	-
{ health infectious public professionals clinicians, typeOf, infectious public professionals clinicians . }	-
{ infectious public professionals clinicians, typeOf, clinicians . }	-
{ disease health infectious public professionals clinicians, typeOf, health infectious public professionals clinicians . }	-
{ infectious public, typeOf, public . }	-
{ infectious public professionals, typeOf, professionals . }	-
{ disease, typeOf, disease infectious clinicians public professionals . }	-
{ concerted effort, typeOf, effort . }	-
{ variety of sources, typeOf, variety . }	-
{ variety of sources, typeOf, sources . }	-
{ ongoing partnership, typeOf, partnership . }	-
{ infectious disease clinicians, typeOf, clinicians . }	-
{ disease infectious clinicians, typeOf, infectious clinicians . }	-
{ infectious clinicians, typeOf, clinicians . }	-
{ infectious clinicians public professionals, typeOf, disease infectious clinicians public professionals . }	-
{ infectious clinicians public professionals, typeOf, health infectious clinicians public professionals . }	-

Simplicity Ratio =12.5/13 = 96%

As calculated above, the simplicity ratio of the generated RDF triples utilizing Boxer output on sentence_4 equals 96% since the object of one of the triples neither is a simple word nor has related triples with the property "typeOf".

Essence Coverage Ratio = (0.5+ 0.5) + (0.5+0.5)/2=100%

As seen in the following RDF triples, they cover the essence of the sentence. The triples contain the main verbs of the sentence including "require" and "include" and their arguments. Therefore, the Boxer essence coverage equals 100%.

Triples with the verb "require":

{ overcome, **require**, concerted effort . }
{ overcome, patient, challenges . }
{ limitations, **require**, concerted effort . }

Triples with the verb "include":

{ variety of sources, **include**, ongoing partnership . }

Words Coverage Ratio= 19/19= 100

The words of the sentence are listed as follows: "1-Overcoming, 2-challenges, 3- limitations, 4- require, 5-concerted, 6-effort, 7- from, 8- variety, 9-sources, 10-including, 11-ongoing, 12-partnership, 13- between, 14-infectious, 15- disease, 16- clinicians, 17- public, 18-health, 19-professionals."

In the generated RDF triples, it is observed that all the above words appeared at least once, therefore, the words coverage of this case study equals 100%.

T2R Results:

The analysis of the Sentence_4 by T2R tool leads to generation of a set of RDF triples as shown in Table 4.15.

Table 4.15 Sentence_4:T2R Output

RDF triples generated using T2R tool	Simplicity Score
{ limitations, require from, partnership-variety of sources . }	0.5
{ challenges, require, concerted effort . }	1
{ challenges, require from, variety of partnership-sources . }	0.5
{ limitations, require from, health professionals public . }	1
{ limitations, require, concerted effort . }	1
{ infectious disease clinicians, typeOf, clinicians . }	-
{ partnership-sources, typeOf, sources . }	-
{ concerted effort, typeOf, effort . }	-
{ partnership-variety, typeOf, variety . }	-
{ health professionals public, typeOf, professionals . }	-
{ clinicians-partnership, typeOf, partnership . }	-
{ ongoing partnership, typeOf, partnership . }	-
{ infectious disease clinicians, typeOf, clinicians . }	-
{ ongoing partnership, typeOf, partnership . }	-
{ clinicians-partnership, typeOf, partnership . }	-
{ concerted effort, typeOf, effort . }	-

Simplicity Ratio= 4/5=80%

As calculated above, the simplicity ratio of the generated RDF triples utilizing T2R output on Sentence_4 equals 85% since the objects of two triples neither are simple words nor have related triples with the property “typeOf”.

Essence Coverage Ratio = (0.5+ 0.5) + (0+0)/2=50%

As seen in the following selected T2R RDF triples, these RDF triples contain the verb “require” but no triple exists with the verb “include". That is, the T2R essence coverage equals 50%.

Triples with the verb “require”:

{ challenges, **require**, concerted effort . }
{ challenges, **require** from, variety of partnership-sources . }
{ limitations, **require**, concerted effort . }
{ limitations, **require** from, partnership-variety of sources . }

Words Coverage Ratio =16/19=84%

Three words including “overcoming”, “include”, and “between” do not exist in the generated T2R RDF triples. Therefore, the words coverage ratio equals 84%.

Confidence level:

Table 4.16 Sentence_4: RDF Triples with the main verb

Tools	main verb (property)	subject	Property	object
Boxer	require	overcome	require	concerted effort
		limitations	require	concerted effort
		overcome	patient	challenges
T2R	require	challenges	require	concerted effort
Boxer	include	variety of sources	include	ongoing partnership
T2R	include	"NO similar nodes"	"NO similar nodes	"NO similar nodes"

As seen in Table 4.16, there is only similarity between triples containing verb "require" in the RDF triples generated by both tools. RDF triples generated by T2R do not include any triples with "include" verb as a node. Therefore, the generated RDF triples have 50% confidence level:

$$\textbf{Confidence Level Ratio} = \frac{\sum \textbf{ the similarity score between the triples including the main verb}}{\textbf{Total number of main verbs}}$$

$$= \textbf{(0.5+0.5) + (0+0)/ 2 =50\%}$$

The values of measures obtained for the Sentence_4 are summarized in Table 4.17.

Table 4.17 Summary of the results for the Sentence_4

Simplicity		Essence coverage		Word coverage		Confidence level
Boxer	**T2R**	**Boxer**	**T2R**	**Boxer**	**T2R**	
96%	**80%**	**100%**	**50%**	**100%**	**84%**	**50%**

Sentence_ 5:

"The basic public health tools of surveillance and epidemiologic investigation helped define the epidemic and led to initial prevention recommendations." [32]

Boxer Results:

The RDF triples generated by Boxer used on the Sentence_5 are shown in Table 4.18.

Table 4.18 Sentence_5: Boxer Output

RDF triples generated using Boxer tool	Simplicity Score
{ basic public health tools of surveillance, define, epidemic . }	1
{ epidemiologic investigation, define, epidemic . }	1
{ lead, agent, basic public health tools . }	1
{ basic public health tools, lead to, initial prevention recommendations . }	1
{ initial, prevention, recommendations . }	1
{ basic public tools, of, surveillance . }	1
{ lead, agent, basic public tools . }	1
{ lead, to, initial recommendations . }	1
{ lead, agent, epidemiologic investigation . }	1
{ basic public tools, define, epidemic . }	1
{ epidemiologic investigation, define, epidemic . }	1
{ basic public health tools of surveillance, typeOf, tools . }	-
{ health basic public tools, typeOf, basic public tools . }	-
{ basic public health tools of surveillance, typeOf, surveillance . }	-
{ basic public health tools of surveillance, typeOf, basic public health tools . }	-
{ basic public, typeOf, public . }	-
{ basic public tools, typeOf, tools . }	-
{ epidemiologic investigation, typeOf, investigation . }	-
{ basic public health tools, typeOf, tools . }	-
{ initial prevention recommendations, typeOf, recommendations . }	-
{ health basic public tools, typeOf, basic public tools . }	-
{ prevention initial recommendations, typeOf, initial recommendations . }	-
{ basic public, typeOf, public . }	-
{ basic public tools, typeOf, tools . }	-
{ initial recommendations, typeOf, recommendations . }	-
{ health, typeOf, health basic public tools . }	-
{ basic public tools, typeOf, health basic public tools . }	-

Simplicity Ratio= 11/11=100%

As seen in Table 4.18, the simplicity score of each triple except for triples with "typeOf" property equals to 1.0. It is due the fact that, for each triple, not only the subject but also the

object of the triple are either simple or have corresponding triples with "typeOf" property that shows the core of the noun phrase.

Essence Coverage Ratio = (0.5+ 0.5) + (0.5+0.5)/2=100%

As seen in the following selected RDF triples, they cover the essence of the sentence. The triples contain the main verbs of the sentence including "define" and "lead" and their arguments. Therefore, the Boxer essence coverage equals 100%.

Triples with the verb "define":

{ epidemiologic investigation, **define**, epidemic . }
{ basic public health tools of surveillance, **define**, epidemic . }

Triples with the verb "lead":

{ basic public health tools, **lead** to, initial prevention recommendations . }

The words of the sentence are listed as follows: "1-basic, 2- public, 3- health, 4- tools, 5- surveillance,6- epidemiologic, 7- investigation, 8-helped, 9-define, 10- epidemic, 11- led to, 12- initial, 13- prevention, 14-recommendations."

Words Coverage Ratio = 13/14=92 %

In the generated Boxer RDF triples, one word "helped" does not exist. Therefore, the words coverage ratio of this case study equals 92 %.

T2R Results:

The analysis of the Sentence_5 by T2R tool leads to generation of a set of RDF triples as shown in Table 4.19.

Table 4.19 Sentence_5: T2R Output

RDF triples generated using T2R tool	Simplicity Score
{ basic public health tools, helped, epidemic . }	1
{ basic public health tools, define, epidemic . }	1
{ basic public health tools of surveillance, helped, epidemic . }	0.5
{ basic public health tools of surveillance, define, epidemic . }	0.5
{ epidemiologic investigation, helped, epidemic . }	1
{ epidemiologic investigation, define, epidemic . }	1
{ basic public health tools, lead to, recommendations initial prevention . }	1
{ basic public health tools of investigation surveillance epidemiologic, helped, epidemic . }	0.5
{ basic public health tools of investigation surveillance epidemiologic, define, epidemic . }	0.5
{ basic public health tools of investigation surveillance epidemiologic, lead to, initial prevention recommendations . }	0.5
{ basic public health tools, typeOf, tools . }	-
{ epidemiologic investigation, typeOf, investigation . }	-
{ recommendations initial prevention, typeOf, recommendations . }	-
{ investigation surveillance epidemiologic, typeOf, investigation . }	-
{ initial prevention recommendations, typeOf, recommendations . }	-

Simplicity Ratio= 7.5/10=75%

As calculated above, the simplicity ratio of the generated RDF triples utilizing T2R output on Sentence_5 equals 75% since the subjects of five triples neither are simple words nor have related triples with the property "typeOf".

Essence Coverage Ratio = (0.5+ 0.5) + (0.5+0.5)/2=100%

As seen in the following selected RDF triples, they cover the essence of the sentence. These triples contain the main verbs of the sentence including "define" and "lead" and their arguments. Therefore the T2R essence coverage equal 100%.

Triples with the verb "define":

{ basic public health tools, **define**, epidemic . }
{ basic public health tools of surveillance, **define**, epidemic . }
{ epidemiologic investigation, **define**, epidemic . }

Triples with the verb "lead":

{ basic public health tools, **lead** to, recommendations initial prevention . }

Words Coverage Ratio = 14/14=100 %

All the above words of the sentence appeared at least once in the generated RDF triples. As a

result, the word coverage ratio of this case study equals 100%.

Confidence level:

Table 4.20 Sentence_5: RDF triples with the main verb(s)

Tools	main verb	subject	Property	Object
Boxer	define	epidemiologic investigation	define	epidemic
		basic public health tools of surveillance	define	epidemic
T2R	define	basic public health tools	define	epidemic
		basic public health tools of surveillance	define	epidemic
		epidemiologic investigation	define	epidemic
Boxer	lead	basic public health tools	lead to	initial prevention recommendations
T2R	lead	basic public health tools	lead to	recommendations initial prevention

As seen in Table 4.20, the triples with "define" and "lead" properties have similar or the same subjects and objects in the RDF triples generated by both tools. That is, the score of 0.5 is given to subjects and the score of 0.5 to objects. The confidence score equals 1.0, which indicates that the generated RDF triples have 100% confidence level:

$$\textbf{Confidence Level Ratio} = \frac{\sum \textbf{the similarity score between the triples including the main verb}}{\textbf{Total number of main verbs}}$$

$$= \textbf{(0.5+0.5) + (0.5+0.5)/ 2 =100\%}$$

The values of measures obtained for the Sentence_5 are summarized in Table 4.21.

Table 4.21 Summary of the results for the Sentence_5

Simplicity		Essence coverage		Word coverage		Confidence level
Boxer	**T2R**	**Boxer**	**T2R**	**Boxer**	**T2R**	
100%	**75%**	**100%**	**100%**	**92%**	**100%**	**100%**

Sentence_6:

"*Even though research in ecology has always had a strong theoretical component, cultural and technical hurdles often hamper direct collaboration between theoreticians and empiricists."* [46]

Boxer Results:

The RDF triples generated by Boxer used on the Sentence_6 are shown in Table4. 22.

Table 4.22 Sentence_6: Boxer Output

RDF triples generated using Boxer tool	Simplicity Score
{ cultural hamper direct collaboration, between, theoreticians . }	1
{ direct collaboration, between, theoreticians . }	1
{ direct collaboration, between, empiricists . }	1
{ research, in, ecology . }	1
{ technical hurdles, hamper often Even, direct collaboration . }	1
{ research, have, strong theoretical component . }	1
{ theoreticians, and, empiricists . }	1
{ direct collaboration, between, theoreticians empiricists . }	0.5
{ research, have always, strong theoretical component . }	1
{ cultural hurdles , often hamper, direct collaboration . }	1
{ cultural hurdles, and , technical hurdles }	1
{ have , when, always . }	1
{ hamper, when, often . }	1
{ cultural hamper direct collaboration, typeOf, collaboration . }	-
{ hamper cultural direct collaboration, typeOf, cultural direct collaboration . }	-
{ cultural direct, typeOf, direct . }	-
{ cultural direct collaboration, typeOf, collaboration . }	-
{ technical hurdles, typeOf, hurdles . }	-
{ direct collaboration, typeOf, collaboration . }	-
{ hamper often, typeOf, hamper . }	-
{ hamper often Even, typeOf, Even . }	-
{ strong theoretical, typeOf, theoretical . }	-
{ strong theoretical component, typeOf, component . }	-
{ have always, typeOf, have . }	-
{ cultural hurdles, typeOf, hurdles}	-
{ technical hurdles, typeOf, hurdles }	-

Simplicity Ratio= 12.5/13=96%

As calculated above, the simplicity ratio of the generated RDF triples utilizing Boxer output on Sentence_6 equals 96% % since the object of one of the triples neither is a simple word nor has a related triple with the property "typeOf".

Essence Coverage Ratio = (0.5+ 0.5) + (0.5+0.5)/2=100%

As seen in the following selected RDF triples, they cover the essence of the sentence. These triples contain the main verbs of the sentence including "have" and "hamper" and their arguments. Therefore, the Boxer essence coverage equals 100%.
Triples with the verb "have":

{ research, **have**, strong theoretical component . }
{ have always, typeOf, have . }
{ have , when, always . }

Triples with the main verb "hamper":

{ technical hurdles, **hamper** often Even, direct collaboration . }
{ cultural hurdles , often **hamper**, direct collaboration . }
{ hamper, when, often . }
{ hamper often, typeOf, hamper . }

Words Coverage Ratio: 17/17=100%

The words of the sentence are listed as follows: "1-research, 2-ecology, 3-always, 4-had, 5-strong, 6-theoretical, 7-component, 8-cultural, 9-technical, 10-hurdles, 11-often, 12-hamper, 13-direct, 14-collaboration, 15-between, 16-theoreticians, 17- empiricists."

In the generated RDF triples, it is observed that all the above words appeared at least once, therefore, the words coverage of this case study equals 100%.

T2R Results:

The analysis of the Sentence_6 by T2R tool leads to generation of a set of RDF triples as shown in Table 4.23.

Table 4.23 Sentence_6: T2R Output

RDF triples generated using T2R tool	Simplicity Score
{ technical hurdles, hamper, often . }	1
{ technical hurdles, hamper, empiricists . }	1
{ technical hurdles, hamper, direct collaboration theoreticians . }	1
{ research, in, ecology . }	1
{ research, have, cultural technical hurdles . }	1
{ have, when, always . }	1
{ research, have, theoretical component strong . }	1
{ research, have, Even . }	1
{ hamper, when, often . }	1
{ cultural technical hurdles, hamper, empiricists . }	1
{ cultural technical hurdles, hamper, direct collaboration theoreticians . }	1
{ cultural technical hurdles, typeOf, hurdles . }	-
{ direct collaboration theoreticians, typeOf, theoreticians . }	-
{ technical hurdles, typeOf, hurdles . }	-
{ direct collaboration theoreticians, typeOf, theoreticians . }	-
{ theoretical component strong, typeOf, component . }	-
{ had-even, typeOf, even . }	-
{ theoretical component strong, typeOf, component . }	-

Simplicity Ratio= 11/11=100%

As calculated above, the simplicity ratio of the generated RDF triples utilizing T2R output on sentence_6 equals 100%. It is due the fact that, for each triple, not only the subject but also the object of the triple are either simple or have corresponding triples with “typeOf” property that shows the core of the noun phrase.

Essence Coverage Ratio = (0.5+ 0.5) + (0.5+0.5)/2=100%

As seen in the following selected RDF triples, these RDF triples cover the essence of the sentence. They contain the main verbs of the sentence including “have” and “hamper” and their arguments. Therefore, the T2R essence coverage equals 100%.
Triples with the verb “have”:

{ research, **have**, theoretical component strong . }
{ research, **have**, cultural technical hurdles . }
{ **have**, when, always . }

Triples with the verb “hamper”:

{ technical hurdles, **hamper**, empiricists . }
{ technical hurdles, **hamper**, direct collaboration theoreticians . }

{ **hamper**, when, often . }

Words Coverage Ratio: 16/17=94%

In the generated T2R RDF triples, one word "between" does not exist. Therefore, the words coverage ratio of this case study equals 94%.

Confidence level:

Table 4.24 Sentence_6: RDF Triples with the main verb

Tools	**main verb (property)**	**subject**	**Property**	**Object**
Boxer	have	research	have	strong theoretical component
		have	when	Always
T2R	have	research	have	theoretical component strong
		research	have	cultural technical hurdles
		have	when	Always
Boxer	hamper	technical hurdles	hamper often Even	direct collaboration
		cultural hurdles	often hamper	direct collaboration
		hamper	when	Often
		hamper often	typeOf	Hamper
T2R	hamper	technical hurdles	hamper	Empiricists
		technical hurdles	hamper	direct collaboration theoreticians
		hamper	when	Often

As seen in Table 4.24, triples with "hamper" and "have" properties have similar or the same subjects and objects in the RDF triples generated by both tools. That is, the score of 0.5 is given to the subjects and the score of 0.5 is given to the objects. The confidence score equals 1, which indicates that the generated RDF triples have 100% confidence level:

$$\textbf{Confidence Level Ratio} = \frac{\sum \textbf{ the similarity score between the triples including the main verb}}{\textbf{Total number of main verbs}}$$

$$= (0.5+0.5) + (0.5+0.5)/\ 2 = 100\%$$

The values of measures obtained for the Sentence_6 are summarized in Table 4.25.

Table 4.25 Summary of the results for the Sentence_6

Simplicity		Essence coverage		Word coverage		Confidence level
Boxer	**T2R**	**Boxer**	**T2R**	**Boxer**	**T2R**	
96%	**100%**	**100%**	**100%**	**100%**	**94%**	**100%**

Sentence_7:

"*Cooperation is needed to prevent the expansion of infections of zoonoses by both human and veterinary medicine."* [47]

Boxer Results:

The RDF triples generated by Boxer used on the Sentence_7 are shown in Table 4.26.

Table 4.26 Sentence_7: Boxer Output

RDF triples generated using Boxer tool	Simplicity Score
{ infections of zoonoses, by, human medicine veterinary . }	1
{ infections of zoonoses, by, medicine . }	1
{ prevent, patient, expansion . }	1
{ prevent, patient, expansion of infections . }	1
{ need, agent, Cooperation . }	1
{ zoonoses, by, human medicine . }	1
{ human medicine, and, veterinary medicine }	1
{ infections, of, zoonoses . }	1
{ expansion, of, infections . }	1
{ Cooperation, need to, expansion . }	1
{ expansion of zoonoses, typeOf, zoonoses . }	-
{ expansion of infections, typeOf, expansion . }	-
{ human medicine, typeOf, medicine . }	-
{ veterinary medicine, typeOf, medicine . }	-
{ human medicine veterinary, typeOf, medicine . }	-
{ medicine, typeOf, human medicine veterinary . }	-
{ infections of zoonoses, typeOf, zoonoses . }	-

Simplicity Ratio= 10/10=100%

As calculated above, the simplicity ratio of the generated RDF triples utilizing Boxer output on Sentence_7 equals 100%. It is due the fact that, for each triple, not only the subject but also the object of the triple are either simple or have corresponding triples with "typeOf" property that shows the core of the noun phrase.

Essence Coverage Ratio = (0+ 0.5)/1=50%

As seen in the following selected RDF triples, the main verb of the sentence "prevent" is not a property in any triples but it can be observed that part of the arguments of the verb appeared in the triples with the property "patient". It indicates that the score of essence coverage is 0.5 and the Boxer essence coverage ratio equals 50%.

Triples with the verb "prevent":

{ **prevent**, patient, expansion . }
{ **prevent**, patient, expansion of infections . }
{ Cooperation, need to, expansion . }

The words of the sentence are listed as follows: "1-Cooperation, 2-needed to, 3-prevent, 4-expansion, 5- infections, 6-zoonoses, 7-human, 8-veterinary, 9-medicine."

Words Coverage Ratio: 9/9=100%

In the generated RDF triples, it is observed that all the above words appeared at least once, therefore, the words coverage of this case study equals 100%.

T2R Results:

The analysis of the Sentence_7 by T2R tool leads to generation of a set of RDF triples as shown in Table 4.27.

Table 4.27 Sentence_7: T2R Output

RDF triples generated using T2R tool	Simplicity Score
{ Cooperation, prevent by, human veterinary medicine . }	1
{ infections, of, zoonoses . }	1
{ Cooperation, prevent, expansion of infections . }	0.5
{ human veterinary medicine, typeOf, medicine . }	-

Simplicity Ratio=2.5/3=83%

As calculated above, the simplicity ratio of the generated RDF triples utilizing T2R output on Sentence_7 equals 83% since the object of one of the triples neither is a simple word nor has a related triple with the property "typeOf".

Essence Coverage Ratio = (0.5+ 0.5)/1=100%

As seen in the following selected RDF triples, they cover the essence of the sentence. These triples contain the main verb of the sentence "prevent" and its arguments. Therefore, the T2R essence coverage equals 100%.

Triples with the verb "prevent":

{ Cooperation, prevent, expansion of infections . }
{ Cooperation, prevent by, human veterinary medicine . }

Words Coverage Ratio: 8/9= 88%

In the generated RDF triples, one word "need" does not exist. As a result, the Words coverage ratio equals 88%.

Confidence level:

Table 4.28 Sentence_7: RDF Triples with the main verb

Tools	**main verb (property)**	**subject**	**property**	**Object**
Boxer	prevent	Prevent	patient	expansion
		Prevent	patient	expansion of infections
T2R	prevent	Cooperation	prevent	expansion of infections
		Cooperation	prevent by	human veterinary medicine

As seen in Table 4.28, since the main verb "prevent" did not appear as a property of the RDF triples generated by Boxer, all the arguments of the verb do not exist in the Boxer RDF triples, and the triples generated by these two tools have only similar objects. Therefore, the generated RDF triples have 50% confidence level:

$$\textbf{Confidence Level Ratio} = \frac{\sum \textbf{the similarity score between the triples including the main verb}}{\textbf{Total number of main verbs}}$$

$$= \textbf{(0.5+0)/1 =50\%}$$

The values of measures obtained for the Sentence_7 are summarized in Table 4.29.

Table 4.29 Summary of the results for the Sentence_7

Simplicity		Essence coverage		Word coverage		Confidence level
Boxer	**T2R**	**Boxer**	**T2R**	**Boxer**	**T2R**	
100%	**83%**	**50%**	**100%**	**100%**	**88%**	**50%**

Sentence_8:

"*The revised International Health Regulations IHR, which requires the Member States of the World Health Organization WHO to develop core capacities to detect, assess, report, and respond to public health threats, is bringing new challenges for national and international surveillance systems."* [48]

Boxer Results:

The RDF triples generated by Boxer used on the Sentence_8 are shown in Table 4.30.

Table 4.30 Sentence_8: Boxer Output

RDF triples generated using Boxer tool	Simplicity Score
{ IHR, Acronym for , InternationalHealthRegulations . }	1
{ InternationalHealthRegulations, require, MemberStates . }	1
{ revise, require, MemberStates . }	1
{ InternationalHealthRegulations, require, MemberStates of WorldHealthOrganization . }	1
{ revise, require, MemberStates of WorldHealthOrganization . }	1
{ MemberStates, type, loc . }	1
{ WorldHealthOrganization, type, org . }	1
{ MemberStates, of, WorldHealthOrganization . }	1
{ detect, patient, assess report . }	0.5
{ respond, to, public threats . }	0.5
{ bring, for, International systems . }	1
{ bring, for, national systems . }	1
{ bring, rel, WorldHealthOrganization . }	1
{ WHO, develop, capacities . }	1
{ WHO, bring, new challenges . }	0.5
{ MemberStates of WorldHealthOrganization, typeOf, WorldHealthOrganization . }	
{ MemberStates of WorldHealthOrganization, typeOf, MemberStates . }	-
{ surveillance, typeOf, surveillance International national systems systems . }	-
{ surveillance, typeOf, surveillance International national systems systems . }	-
{ Health, typeOf, Health public threats . }	-
{ core, typeOf, core capacities . }	-
{ International systems, typeOf, systems . }	-
{ national systems, typeOf, systems . }	-
{ public threats, typeOf, Health public threats . }	-
{ capacities, typeOf, core capacities . }	-

Simplicity Ratio= 13.5/15= 90%

As calculated above, the simplicity ratio of the generated RDF triples utilizing Boxer output on Sentence_8 equals 90% since the objects of three triples neither are simple words nor have related triples with the property “typeOf”.

Essence Coverage Ratio = (0.5+ 0.5) + (0.5+0.5) + (0+0.5)/3=83%

As seen in the following selected RDF triples, they cover the essence of the sentence. These triples contain the main verb of the sentence including “require” and “develop” and their arguments, but for the triple which has the verb “develop” as its property, the subject of the triple is “WHO” which is not the correct subject. Accordingly, the Boxer essence coverage ratio equals

83%.

Triples with the verb “require”:

{ InternationalHealthRegulations, **require**, MemberStates . }
{ revise, **require**, MemberStates . }
{ InternationalHealthRegulations, **require**, MemberStates of WorldHealthOrganization . }

Triples with the verb “develop”:

{WHO, **develop**, capacities. }

Triples with the verb “bring”:

*{WHO, **bring** ,new challenges.}*

Words Coverage Ratio = 29/29= 100%

The words of the sentence are listed as follows: 1-revised, 2- International, 3-Health,4-Regulations,5- IHR, 6-requires, 7- Member, 8- States, 9- World, 10-Health, 11- Organization, 12-WHO, 13-develop, 14- core, 15-capacities, 16-detect, 17-assess, 18-report, 19- respond, 20-public, 21-health, 22-threats, 23-bringing , 24-new, 25-challenges, 26-national, 27-international, 28-surveillance, 29-systems.

In the generated RDF triples, it is observed that all the above words appeared at least once, therefore, the words coverage of this case study equals 100%.

T2R Results:

The analysis of the Sentence_8 by T2R tool leads to generation of a set of RDF triples as shown in Table 4.31.

Table 4.31 Sentence_8: T2R Output

RDF triples generated using T2R tool	Simplicity Score
{ WorldHealthOrganization, is, Loc . }	1
{ WorldHealthOrganization, is, Misc . }	1
{ MemberStates, is, Misc . }	1
{ MemberStates of WorldHealthOrganization, is, Misc . }	0.5
{ revised InternationalHealthRegulations, is, Org . }	1
{ revised InternationalHealthRegulations, require, core capacities . }	1
{ anonymous, develop, core capacities . }	1
{ respond, to, public health threats . }	1
{ MemberStates, is, Org . }	1
{ MemberStates, is, respond-reportand . }	1
{ revised InternationalHealthRegulations, is, Loc . }	1
{ new challenges, for, surveillance systems national international . }	1
{ revised InternationalHealthRegulations, bring, new challenges . }	1
{ anonymous, bring, new challenges . }	1
{ revised InternationalHealthRegulations, require, MemberStates of WorldHealthOrganization . }	0.5
{ MemberStates of WorldHealthOrganization, is, Org . }	0.5
{ MemberStates of WorldHealthOrganization, is, respond-reportand . }	0.5
{ InternationalHealthRegulations, Acronym for, IHR . }	1
{ core capacities, typeOf, capacities . }	-
{ respond-reportand, typeOf, reportand . }	-
{ public health threats, typeOf, threats . }	-
{ new challenges, typeOf, challenges . }	-
{ surveillance systems national international, typeOf, systems . }	-
{ revised InternationalHealthRegulations, typeOf, InternationalHealthRegulations . }	-

Simplicity Ratio= 16/18=88%

As calculated above, the simplicity ratio of the generated RDF triples utilizing T2R output on Sentence_8 equals 88% since the subjects of three triples and object of one triple neither are simple words nor have related triples with the property “typeOf”.

Essence Coverage Ratio = (0.5+0.5) + (0+0.5) + (0.5+0.5)/3=83%

As seen in the following selected RDF triples, they cover the essence of the sentence. These triples contain the main verb of the sentence including “require” and “bring” and their arguments, but for the triple which has the verb “develop” as its property, the subject of the triple is “anonymous” which indicates that this triple does not contain the correct subject of the verb. That is, the T2R essence coverage ratio equals 83%.

Triples with the verb “require ”:

{ revised InternationalHealthRegulations, **require**, MemberStates of WorldHealthOrganization .
}

Triples with the verb “develop”:

{ anonymous, **develop**, core capacities . }

Triples with the verb “bring”:

{ revised InternationalHealthRegulations, **bring**, new challenges . **}**

Words Coverage Ratio =25/29= 86%

Four words including “WHO”, “detect”, “assess”, and “report” do not exist in the generated T2R RDF triples. Therefore, the Words coverage ratio equals 86%.

Confidence-level:

Table 4.32 Sentence_8: RDF Triples with the main verb

Tools	**main verb (property)**	**subject**	**Property**	**Object**
Boxer	require	InternationalHealthRegulations	require	MemberStates
		Revise	require	MemberStates
		InternationalHealthRegulations	require	MemberStates of WorldHealthOrganization
T2R	require	revised InternationalHealthRegulations	require	MemberStates of WorldHealthOrganization
Boxer	develop	WHO	develop	Capacities
T2R	develop	anonymous	develop	core capacities
Boxer	bring	WHO	bring	new challenges
T2R	bring	revised InternationalHealthRegulations	bring	new challenges

As seen in Table 4.32, there is a similarity between the nodes of the triple with the verb “require” as a property. That is, the confidence score for this main verb equals 1.0 but the subjects of the triples with the property “develop” are not similar since T2R triples cannot detect

the subject of the verb "develop. Therefore, the confidence score for this main verb equals 0.5. The subjects of triples with property "bring" are not the same. Therefore the confidence score of this verb equals 0.5. As a result, the generated RDF triples have 66% confidence level:

$$\textbf{Confidence Level Ratio} = \frac{\sum \textbf{the similarity score between the triples including the main verb}}{\textbf{Total number of main verbs}}$$

$$= (0.5+0.5) + (0.5+0) + (0.5+0)/3 = 66\%$$

The values of measures obtained for the Sentence_8 are summarized in Table 4.33.

Table 4.33 Summary of the results for the Sentence_8

Simplicity		Essence coverage		Word coverage		Confidence level
Boxer	T2R	Boxer	T2R	Boxer	T2R	
90%	88%	83%	83%	100%	86%	66%

Sentence_9:

"The use of the continental surveillance system is the main strategy for achieving the eradication of FMD in South America." [49]

Boxer Results:

The RDF triples generated by Boxer used on the Sentence_9 are shown in Table 4.34.

Table 4.34 Sentence_9: Boxer Output

RDF triples generated using Boxer tool	Simplicity Score
{ achieve, in, SouthAmerica . }	1
{ use of continental surveillance system, is, main strategy . }	1
{ main strategy, achieve, eradication . }	1
{ main strategy, achieve, eradication of Fmd . }	1
{ fmd, type, org . }	1
{ SouthAmerica, type, loc . }	1
{ use, is, main strategy . }	1
{ use, of, continental system . }	1
{ eradication, of, FMD . }	1
{ surveillance, typeOf, surveillance continental system . }	-
{ continental system, typeOf, surveillance continental system . }	-
{ continental surveillance system, typeOf, system . }	-
{ surveillance continental system, typeOf, continental system . }	-
{ use of continental surveillance system, typeOf, continental surveillance system . }	-
{ use of continental surveillance system, typeOf, use . }	-
{ eradication of Fmd, typeOf, Fmd . }	-
{ eradication of Fmd, typeOf, eradication . }	-
{ continental system, typeOf, system . }	-
{ main strategy, typeOf, strategy . }	-

Simplicity Ratio= 9/9=100%

As calculated above, the simplicity ratio of the generated RDF triples utilizing Boxer output on Sentence_9 equals 100%. It is due the fact that, for each triple, not only the subject but also the object of the triple are either simple or have corresponding triples with "typeOf" property that shows the core of the noun phrase.

Essence Coverage Ratio = (0.5+0.5) + (0.5+0.5)/2=100%

As seen in the following selected RDF triples, they cover the essence of the sentence. These triples contain the main verb of the sentence including "is" and "achieve" and their arguments. Therefore, the Boxer essence coverage for equals 100%.

Triples with the main verb "is":

{ use of continental surveillance system, **is**, main strategy . }

Triples with the main verb “achieve”:

{ main strategy, **achieve**, eradication of Fmd . }

Words Coverage Ratio=10/10=100%

The words of the sentence are listed as follows: “1-use, 2- continental, 3- surveillance, 4-system, 5- main, 6-strategy, 7-achieving, 8-eradication, 9-FMD, 10- South America.”

In the generated RDF triples, it is observed that all the above words appeared at least once, therefore, the words coverage of this case study equals 100%.

T2R Results:

The analysis of the Sentence_9 by T2R tool leads to generation of a set of RDF triples as shown in Table 4.35.

Table 4.35 Sentence_9: T2R Output

RDF triples generated using T2R tool	Simplicity Score
{ achieve, in, SouthAmerica . }	1
{ SouthAmerica, is, Loc . }	1
{ FMD, is, Org . }	1
{ use, is, main strategy . }	1
{ use of continental surveillance system, is, main strategy . }	1
{ anonymous, achieve, eradication of FMD . }	1
{ eradication of FMD, typeOf, FMD . }	-
{ use of continental surveillance system, typeOf, continental surveillance system . }	-
{ main strategy, typeOf, strategy . }	-
{ continental surveillance system, typeOf, system . }	-

Simplicity Ratio= 6/6=100%

As calculated above, the simplicity ratio of the generated RDF triples utilizing T2R output on Sentence_9 equals 100%.

Essence coverage ratio= (0.5+0.5) + (0+0.5)/2=0.75

As seen in the following selected RDF triples, they cover the essence of the sentence with the verb “is” and its arguments but triples with the other verb of the sentence “achieve” does not have clear subject (anonymous). Therefore, this indicates that the score of essence coverage for this verb is 0.5. As a result, the T2R essence coverage equals 75%.

Triples with the main verb “is”:

{ use of continental surveillance system, **is**, main strategy . }

Triples with the main verb “achieve ”:

{ anonymous, **achieve**, eradication of FMD . }

Words Coverage Ratio=10/10=100%

The word coverage ratio equals 100% since all the listed words appeared at least once in the generated RDF triples.

Confidence-level:

Table 4.36 Sentence_9: RDF Triples with the main verb

Tools	**main verb (property)**	**subject**	**Property**	**Object**
Boxer	is	use of continental surveillance system,	is	main strategy
T2R	is	use of continental surveillance system, ,	is	main strategy
Boxer	achieve	main strategy	achieve	eradication of Fmd
T2R	achieve	anonymous	achieve	eradication of FMD

As seen in Table 4.36, there is similarity between the nodes of the triples with property “is”. That is, the confidence score for this main verb equals 1.0, but the subjects of the triples with “achieve” property are not similar since T2R triples cannot detect the subject of the “achieve” verb, therefore the confidence level score of this verb equals 0.5. As a result, the generated RDF triples have 75% confidence level:

$$\textbf{Confidence Level Ratio} = \frac{\sum \textbf{ the similarity score between the triples including the main verb}}{\textbf{Total number of main verbs}}$$

$$= \textbf{(0.5+0.5) + (0.5+0)/2=75\%}$$

The values of measures obtained for the Sentence_9 are summarized in Table 4.37.

Table 4.37 Summary of the results for the Sentence_9

Simplicity		Essence coverage		Word coverage		Confidence level
Boxer	**T2R**	**Boxer**	**T2R**	**Boxer**	**T2R**	
100%	**100%**	**100%**	**75%**	**100%**	**100%**	**75%**

Sentence_ 10:

"The Armed Forces Health Surveillance Center, Global Emerging Infections Surveillance and Response System (AFHSC-GEIS) has the mission of performing surveillance for emerging infectious diseases that could affect the United States (U.S.) military." [50]

Boxer Results:

The RDF triples generated by Boxer used on the Sentence_10 are shown in Table 4.38.

Table 4.38 Sentence_10: Boxer Output

RDF triples generated using Boxer tool	Simplicity Score
{ perform, for, emerge infectious diseases . }	1
{ mission, perform, Surveillance . }	1
{ ArmedForcesHealthSurveillanceCenterGlobalEmergingInfectionsSurveillance, have, mission . }	0.5
{ ResponseSystem, have, mission . }	0.5
{ emerge infectious diseases, affect, military . }	1
{ US, Acronym for , UnitedStates . }	1
{ AFHSC, Acronym for , ForcesHealthSurveillanceSystemCenter . }	1
{ GEIS, Acronym for , EmergingInfectionsSurveillanceSystem . }	1
{ military, typeOf, military . }	
{ emerge infectious diseases, typeOf, diseases . }	

Simplicity Ratio= 7/8=87%

As calculated above, the simplicity ratio of the generated RDF triples utilizing Boxer output on Sentence_10 equals 87% since the subjects of two triples neither are simple words nor have related triples with the property "typeOf".

Essence coverage ratio= (0.5+0.5) + (0.5+0.5)+(0.5+0.5)/3=100%

As seen in the following selected RDF triples, they cover the essence of the sentence. These triples contain the main verb of the sentence including “have”, ”perform”, and “affect” and their arguments. Therefore, the Boxer essence coverage equals 100%.

Triples with the main verb “have”:

{ ResponseSystem, **have**, mission . }
{ ArmedForcesHealthSurveillanceCenterGlobalEmergingInfectionsSurveillance, **have**, mission . }

Triples with the main verb “perform”:

{ mission, **perform**, Surveillance . }

Triples with the main verb “affect”:

{ emerge infectious diseases, **affect**, military . }

Words Coverage Ratio=23/23=100%

The words of the sentence are listed as follows: “1-Armed , 2-Forces , 3-Health, 4-Surveillance, 5-Center, 6-Global, 7-Emerging, 8-Infections, 9-Surveillance , 10-Response, 11-System, 12-(AFHSC-GEIS), 13-has , 14-mission , 15-performing , 16-surveillance , 17-emerging, 18-infectious, 19-diseases , 20-affect , 21-United States, 22-(U.S.) , 23-military.”

In the generated RDF triples, it is observed that all the above words appeared at least once, therefore, the words coverage of this case study equals 100%.

T2R Results:

The analysis of the Sentence_10 by T2R tool leads to generation of a set of RDF triples as shown in Table 4.39.

Table 4.39 Sentence_10: T2R Output

RDF triples generate using T2R tool	Simplicity Score
{ ArmedForcesHealthSurveillanceCenterGlobalEmergingInfectionsSurveillance, have, surveillance infectious diseases . }	0.5
{ ArmedForcesHealthSurveillanceCenterGlobalEmergingInfectionsSurveillance, is, Org . }	0.5
{ ArmedForcesHealthSurveillanceCenterGlobalEmergingInfectionsSurveillance, have, mission . }	0.5
{ anonymous, perform, surveillance infectious diseases . }	1
{ surveillance, for, surveillance infectious diseases . }	1
{ ResponseSystem, have, mission . }	0.5
{ ResponseSystem, have, surveillance infectious diseases . }	0.5
{ ResponseSystem, is, Org . }	0.5
{ surveillance infectious diseases, typeOf, diseases . }	-
{ surveillance infectious diseases, typeOf, diseases . }	-
{ emerging infectious diseases, typeOf, diseases . }	-

Simplicity Ratio= 5/8=62%

As calculated above, the simplicity ratio of the generated RDF triples utilizing T2R output on Sentence_10 equals 62%since the subjects of six triples neither are simple words nor have related triples with the property “typeOf”.

Essence Coverage Ratio = (0.5+0.5) + (0+0.5) + (0+0)/3= 50%

As seen in the following selected RDF triples, they cover the essence of the sentence with “have” verbs and its arguments but the triple with “perform” verb does not have clear subject (anonymous). Therefore, this indicates that the score of essence coverage for this verb is 0.5. In this case study, T2R cannot provide triples having “affect” verb as a node. The T2R essence coverage ratio equals 50%.

Triples with the main verb “have”:

{ ArmedForcesHealthSurveillanceCenterGlobalEmergingInfectionsSurveillance, **have**, mission . }

{ ResponseSystem, **have**, mission . }

Triples with the main verb “perform”:

{ anonymous, **perform**, surveillance infectious diseases . }

Words Coverage Ratio=19/23=82%

All the words of the sentences except four words including "affect", "United States", "U.S.", "military" appeared at least once in the generated RDF triples. As a result, the word coverage ratio equals 82%.

Confidence-level:

Table 4.40 Sentence_10: RDF Triples with the main verb

Tools	**main verb (property)**	**subject**	**Property**	**Object**
Boxer	have	ResponseSystem	have	mission
		ArmedForcesHealthSurveillanceCenterGlobalEmergingInfectionsSurveillance, have, mission	have	mission
T2R	have	ArmedForcesHealthSurveillanceCenterGlobalEmergingInfectionsSurveillance	have	mission
		ResponseSystem	have	mission
Boxer	perform	Mission	perform	surveillance
T2R	perform	Anonymous	perform	surveillance infectious diseases
Boxer	affect	emerge infectious diseases	affect	military
T2R	affect	"NO similar nodes"	"NO similar nodes"	"NO similar nodes

As seen in Table 4.40, there is a similarity between the nodes of the triple with "have" verb as a property. As a result, the confidence score for this main verb equals 1.0, but the subjects of the triples with "perform" property are not similar since T2R triples cannot detect the subject of the "perform" verb, therefore the confidence score of this verb equals 0.5. Moreover, there is no equivalent or similar triples for the verb "affect" in produced T2R triples. Therefore the confidence score for this verb is 0.0. As a result, the generated RDF triples have 50% confidence level:

$$\textbf{Confidence Level Ratio} = \frac{\sum \textbf{the similarity score between the triples including the main verb}}{\textbf{Total number of main verbs}}$$

$$= (0.5+0.5) + (0.5+0) + (0+0) = 1.5/3 = 50\%$$

The values of measures obtained for the Sentence_10 are summarized in Table 4.41.

Table 4.41 Summary of the results for the Sentence_10

Simplicity		Essence coverage		Word coverage		Confidence level
Boxer	**T2R**	**Boxer**	**T2R**	**Boxer**	**T2R**	
87%	**62%**	**100%**	**50%**	**100%**	**82%**	**50%**

4.42 Summary of the results for the case study A

Sentence No.	Simplicity		Essence coverage		Word coverage		Confidence level
	Boxer	**T2R**	**Boxer**	**T2R**	**Boxer**	**T2R**	
Sentence_1	100%	100%	100%	100%	100%	100%	100%
Sentence_2	100%	100%	100%	50%	100%	85%	50%
Sentence_3	95%	100%	100%	100%	100%	100%	100%
Sentence_4	96%	80%	100%	50%	100%	84%	50%
Sentence_5	100%	75%	100%	100%	92%	100%	100%
Sentence_6	96%	100%	100%	100%	100%	94%	100%
Sentence_7	100%	83%	50%	100%	100%	88%	50%
Sentence_8	90%	88%	83%	83%	100%	86%	66%
Sentence_9	100%	100%	100%	75%	100%	100%	75%
Sentence_10	87%	62%	100%	50%	100%	82%	50%
Average	**96.4%**	**88.8%**	**93.3%**	**80.8%**	**99.2%**	**91.9%**	**74.1%**
Standard Deviation	**4.6%**	**13.5%**	**16.1%**	**22.9%**	**2.5%**	**7.6%**	**23.7%**

4.2.2 Case Study B

Sentence_ 1:
*"These may be attributable to the emergence of a hypervirulent strain of C. difficile that produces increased levels of toxins A and B, as well as an extra toxin known as "binary toxin." This previously uncommon strain has become epidemic, coincident with **its** development of increased resistance to fluoroquinolones, the use of which is increasingly associated with CDAD outbreaks."* [45]

A part of RDF triples generated using Boxer:

{ development, of, **neuter** . } → { development, of, **previously uncommon strain** . }
{ development of increase resistance, to, fluoroquinolones . }
{ become, with, development . }
{ uncommon strain previously, become, epidemic coincident . }
{ become, with, development of increase resistance . }
{ development of **previously uncommon strain**, typeOf**, previously uncommon strain**. }
{ development of **previously uncommon strain** , typeOf, development . }
{ development of increase resistance, typeOf, increase resistance . }
{ development of increase resistance, typeOf, development . }

In the above selected triples, the word "neuter", which refers to the pronoun "its" in the sentence, is replaced with its reference (i.e. "previously uncommon strain" in the triple "{development, of, neuter.}"). In addition, any occurrence of "neuter" is replaced with its reference as shown in bold in the triples.

A part of RDF triples generated using T2R:

{ coincident, with, development of increased resistance . }
{ previously uncommon strain epidemic, coincident with, development of increased resistance . }
{ development of increased resistance, typeOf, increased resistance . }
{ development resistance, typeOf, resistance . }
{ development resistance, to, fluoroquinolones . }
{development resistance, possession, its. }→{ development resistance, of , **previously uncommon strain**. }

As shown in the above triples, in the triple "{development resistance, possession, its.}", instead of the property "possession" and instead of the object "its", the property "of" and the object "previously uncommon strain" are respectively replaced.

As seen in the above triples generated by utilizing Boxer and T2R tools, the reference of the pronoun " its" was correctly resolved by Stanford co-reference chain tool, and it was replaced correctly in the RDF triples. Therefore, the clarity ratio of these RDF triples equals 100%.

Sentence_2:

*"In addition, a major byproduct of the network, and now one of **its** strongest assets, has been the growth of partnerships between ISTM, Centers for Disease Control and Prevention and health-care providers around the world, as well as other medical societies, government, and private organizations."* [51]

A part of RDF triples generated using Boxer:

{one thing, is, growth.}→{ one **major byproduct of the network**, is, growth . }
{one **major byproduct of the network**, of, assets, . }
{ one of assets, is, growth of partnerships . }
{ one of assets, typeOf, assets . }
{ one of assets, typeOf, one . }

In the above selected triples, the word "thing", which refers to the pronoun "its" in the sentences, is replaced with its reference (i.e. "major byproduct of the network" in the triple "{one thing, is, growth.}"). Therefore, any occurrence of "thing" is replaced with its reference as shown in bold in the triples.

A part of RDF triples generated using T2R:

{strongest assets, possession, its.}→{ strongest assets, of, **major byproduct of the network**. }
{ strongest assets, typeOf, assets . }
{ major byproduct of strongest assets, typeOf, strongest assets . }
{ now one-networkand of strongest assets, typeOf, strongest assets . }

In the above selected RDF triples, the triple "{strongest assets, possession, its.}" is changed to triple "{strongest assets, of, major byproduct of the network.}. The property "possession" is changed to the property "of" and the pronoun "its" is replaced with its reference which is "major byproduct of the network".

As seen in the above triples generated by utilizing Boxer and T2R tools, the reference of the pronoun " its" was correctly resolved by Stanford co-reference chain tool, and it was replaced in the RDF triples correctly. Therefore, the clarity ratio of these RDF triples equals 100%.

Sentence_3:

*"Infectious disease ecology has recently raised **its** public profile beyond the scientific community due to the major threats that wildlife infections pose to biological conservation, animal welfare, human health and food security."* [46]

A part of RDF triples generated using C&C-boxer:

{public profile, of, **neuter**. }→**{ public profile, of, recently. }**
{ public profile, beyond, scientific community . }
{ public profile, typeOf, profile . }
{ Infectious ecology, raise due recently, public profile . }

In the above selected triples, the object "neuter" is replaced with the reference which in this case is recognized with the word "recently". Therefore, triple "{public profile, of, neuter. }" is changed to the triple "{ public profile, of, recently . }".

A part of RDF triples generated using T2R:

{ Infectious disease ecology, raise, public profile . }
{ public profile, typeOf, profile . }
{ public profile, possession, **its** . }→**{ public profile, of, recently . }**

In the above selected RDF triples, in the triple "{public profile, possession, its. }" the object "possession" is replaced with the property "of" and the object "its" is replaced with the reference recognized by Stanford Co-reference chain tool which is the word "recently":

For this sentence, the reference of the pronoun "its" was not recognized correctly, since the reference of the pronounce "its" is the noun phrase "Infectious disease ecology". Therefore, the clarity ratio of these RDF triples equals 0%.

Sentence_4:

*"ASM LabCap utilizes ASM's vast resources and **its** membership's expertise-40,000 microbiologists worldwide-to strengthen clinical and public health laboratory systems in low and low-middle income countries."* [52]

A part of RDF triples generated using Boxer:

{ membership microbiologists, of, **neuter** . }→ { membership microbiologists, of, **ASMs** . }
{ strengthen, patient, health laboratory systems . }
{ AsmLabcap, utilize, **ASMs** . }

{ membership microbiologists, strengthen, health laboratory systems . }
{ membership microbiologists of **ASMs**, typeOf, **ASMs** . }
{membership microbiologists of **ASMs**, typeOf, membership microbiologists. }

As seen in the above selected triples, the object "neuter" is replaced with the reference which is "ASMs" and any other occurrence of the word "neuter" related to this reference is replaced with the word "ASMs" as shown in bold.

A part of RDF triples generated using T2R:

{ memberships microbiologists, possession, **its** .}→{ memberships microbiologists, **of**, **ASMs** }
{ ASMLabCap, is, User . }
{ ASMs, is, Entity-utilized . }
{ anonymous, utilize, memberships microbiologists . }
{ memberships microbiologists, typeOf, microbiologists . }

In the above selected RDF triples the pronounce "its" is replaced with the reference which is "ASMs".

As seen in the above RDF triples, the reference of the pronoun is correctly recognized by the Stanford co-reference chain tool and is replaced in the related triples. Therefore, the clarity ratio of these RDF triples equals 100%.

Sentence_5:

"The original designation of laboratories as Levels A, B, C, and D was revised to Sentinel, Reference, and Federal laboratories. ***They*** *now function as an integrated network, with the major goal being to ensure that the nation's public health and private sector laboratories, along with other select laboratories, are prepared and equipped to respond to a biological or chemical act of terrorism in an appropriate and integrated manner."* [53]

A part of RDF triples generated using Boxer:

{ function, agent, thing . }
{ function, as, integrated network . }
{ function, with, major goal . }

A part of RDF triples produced using T2R:

{ they, function, now . }
{ they, function, major goal . }
{ they, function, integrated network being . }

As seen in the above triples, the Stanford co-reference chain resolver cannot resolve the reference of the pronoun in this sentence (i.e. “they”). Consequently, the word “thing”, which represents the pronoun in the Boxer output, and “they” in T2R triples are not replaced with the reference. Therefore, the clarity ratio of these RDF triples equals 0%.

Table 4.43 Summary of the results for the case study B:

Sentence No.	Clarity ratio
Sentence_1	100%
Sentence_2	100%
Sentence _3	0%
Sentence_ 4	100%
Sentence_ 5	0%
Average=60%	

4.2.3 Case study C

In this study, we proposed specific patterns for detecting acronyms in given sentences. All sentences are checked for the acronyms before they are fed to the Boxer and T2R tools. Based on the results, dedicated triples with the property “acronym for” are generated.

Sentence_1:

“There have been recent, marked increases in the incidence and severity of Clostridium difficile-associated disease (CDAD).” [45]

Generated RDF triple with the property “Acronym for”:

{ CDAD, Acronym for , ClostridiumDifficileAssociatedDisease. }

In the above generated RDF triples, it is seen that the acronym and its related words are recognized properly. Therefore, the clarity ratio of these RDF triples equals 100%.

Sentence_2:

“The characteristics of patients with suspected severe acute respiratory syndrome (SARS) hospitalized in a hospital in Paris.” [35]

Generated RDF triple with the property "Acronym for":

{ SARS, Acronym for , severe acute respiratory syndrome . }

In the above generated RDF triples, it is seen that the acronym and its related words are recognized properly. Therefore, the clarity ratio of these RDF triples equals 100%.

Sentence_3:

"Field epidemiology and laboratory training programs (FELTPs) have made significant contributions to public health systems for more than 10 years by producing highly skilled field epidemiologists." [54]

Generated RDF triple with the property "Acronym for":

{ FELTPs, Acronym for , Field epidemiology and laboratory training programs. }

In the above generated RDF triples, it is seen that the acronym and its corresponding words are recognized properly. Therefore, the clarity ratio of these RDF triples equals 100%.

Sentence_4:

"The Laboratory Response Network (LRN) was established in 1999 in response to the worldwide concern for the potential use of biological or chemical agents in the commission of acts of terrorism." [53]

Generated RDF triple with the property "Acronym for":

{ LRN, Acronym for, LaboratoryResponseNetwork . }

In the above generated RDF triples, it is seen that the acronym and its related words are recognized properly. Therefore, the clarity ratio of these RDF triples equals 100%.

Sentence_5:

"Emerging infectious disease outbreaks were identified by the global network and included a wide spectrum of support activities in collaboration with host country partners, several of which were in direct support of the World Health Organization's (WHO) International Health Regulations (IHR) (2005)." [55]

Generated RDF triple with the property "Acronym for":

{ WHO, Acronym for, WorldHealthOrganizations . }

{ IHR, Acronym for, InternationalHealthRegulations . }

In the above generated RDF triples, it is seen that the acronyms and their related words are recognized properly. Therefore, the clarity ratio of these RDF triples equals 100%.

As seen in the above results, for all those sentences that have acronyms, the triple with the property "Acronym for" is generated. Therefore, the average score of the clarity ratio equals 100% as summarized in Table 4.43:

Table 4.44 Summary of the results for the case study C

Sentence No.	Clarity ratio
Sentence_1	100%
Sentence_2	100%
Sentence _3	100%
Sentence_ 4	100%
Sentence_ 5	100%
Average=100%	

4.3 Discussion

Throughout the evaluation process, we want to mimic human ways of evaluating RDF triples generated based on texts. In this process we attempt to simulate a number of aspects which human can see as important concepts shown in the RDF triples. Therefore, we propose three different criteria:

1) Simplicity: it reflects a human user desire for as simple triples as possible which means clear or well-organized nodes.

2) Coverage: it emphasises the need for representation of full content of the text. We defined essence and word coverage criteria. The essence coverage means that essence of the sentence has

to be included in the RDF triples. The word coverage represents the overall need to include all important words of the text.

3) Clarity: It is defined to recognize and evaluate the ambiguities that may exist in RDF triples.

In this section, we have provided the results of application of these criteria on the number of randomly selected sentences. The presented results indicate that the proposed criteria seem to be a good approximation of human evaluation.

The purpose of the forth proposed criterion, confidence level, is to identify the most trustworthy triples obtained using multiple NLP tools. When the triples have high confidence level, it means that there is "agreement" between tools and these triples constitute the true representation of the text.

5. Conclusion and Future work

5.1. Conclusion

The Semantic Web – and particularly the Resource Description Framework (RDF) as a new, different way of representing information – seems to provide quite a different way of looking at data and information stored on the web. This RDF-based form of expressing data means that all pieces of data are connected together as a huge linked dataset of entities and relations. This web of data has one more advantage – it is machine-readable.

The current utilization of the web is an indication that textual documents and a simple text represent and will represent majority of content of Internet. This form of information is the most natural for the users. Therefore, in order to achieve, at least partially, the RDF-based web there have to be methods and tools to convert text into RDF triples.

In the presented work, the main goal is to automatically convert text into RDF triples. We want to make this process as through as possible creating RDF triples that truly represent a text being converted. The approach proposed here is based on a simple idea that there is no single tool that is able to work well and ensure good translation of text into triples. Therefore, the idea is to use a few tools, analyze results obtained from these tools, and finally combine the results and translate them into RDF triples.

On the onset of the proposed approach, we take a text document and preprocess it sentence by sentence. The two stages of preprocessing prepare sentences for parsing. The sentences are become simpler and "clear" of complex and difficult to parse parts. Next, the parsers analyze the sentences, and disambiguation (entity recognition) is performed using Alchemy. The results are run thru a number of pattern recognition procedures and algorithms – a set of RDF triples is crated.

For evaluation, we propose and define a number of criteria: simplicity, coverage, word coverage, clearness, and confidence level. All of those criteria are applied to RDF triples generated based on several randomly selected sentences.

Overall, the proposed method leads to a number of contributions that can be summarized with the following points:

- advanced pre-processing leading to simplification of sentences: to facilitate analyzing sentences by different parsers/tools, an advance pre-processing method that converts complex sentences to the simple ones has been proposed and applied.
- application of multiple parses: since no single tool exists to generate complete results from input sentences, we have utilized different natural language processing tools to obtain more comprehensive results.
- application of Alchemy for entity recognition: by using AlchemyAPI, we are able to obtain more elaborated information or definition of any named entities on the web, and disambiguate those named entities.
- fusion of parsing and Alchemy results for creating RDF triples: we combined the results of natural language processing tools to generate RDF triples.
- proposing evolution criteria for RDF generation process: we proposed a new method for evaluating generated RDF triples without involving human beings through the criteria defined in this study.

5.2 Future Works

Directions for potential future works include:

Fusion with PropBank: The proposed system can benefit from merging it with a RDF version of PropBank. This would provide additional semantics regarding relations and resources/entities extracted from sentences and involved in relations with identified verbs. The obtained sets of RDF triples would represent an attempt to a grammar learning system.

Text Visualization: Once, textual documents are translated into RDF data they can be visualized as RDF graphs. Fusion of multiple RDF triples would lead to visualization of text. Addition of some statistical analysis would provide an interesting way of visual analysis of text. This can be combined with a simple SPARQL query engines to create interactive text querying system.

Text Essence Identification: Another interesting and challenging area of further activities is related to identification of the most essential elements of a text. This would represent an extension and further development of a proposed concept of confidence levels. Such a process would involve ranking the relations extracted from that text and identifying the most dominating ones representing the most important aspect of the text. This could be combined with text visualization. All this would lead to an interesting system of analyzing textual documents, and building data models representing these documents.

Bibliography

[1] B. Nguyen and B. Sameer, A Review of Relation Extraction, Literature review for Language and Statistics II , 2007.

[2] D. Fang, C. Yueguo and D. Xiaoyong, “Linking Entities in Unstructured Texts with RDF Knowledge Bases,” in 15th Asia-Pacific Web Conference, APWeb , Sydney, Australi, 2013.

[3] The Gate project team, "GATE.ac.uk - sale/talks/stupidpoint," [Online]. Available: gate.ac.uk/sale/talks/stupidpoint/diana-fb.ppt. [Accessed 01 06 2014].

[4] C. Siefkes, "Learning to Extract Information for the Semantic Web," in Berliner XML Tage 2003, Berlin, Germany, 2003.

[5] H. Pascal, K. Markus and R. Sebastian, Foundations of Semantic Web Technologies, Chapman and Hall/CRC , 2009.

[6] I. Demiral, "What is the RDFa," PBworks, 2008. [Online]. Available: http://ibrahimdemirall.pbworks.com/w/page/52586364/What%20is%20the%20RDFa. [Accessed 01 06 2014].

[7] S. Pandit, Ontology-guided extraction of structured information from unstructured text:Identifying and capturing complex relationships, Ames, Iowa: M.Sc. thesis, Iowa State University, 2010.

[8] I. AlchemyAPI, “AlchemyAPI,” AlchemyAPI, Inc, 2014. [Online]. Available: http://www.alchemyapi.com/. [Accessed 01 05 2014].

[9] J. Curran, S. Clark and J. Bos, “Linguistically Motivated Large-Scale NLP with C&C and Boxer,” in Proceedings of the ACL 2007 Demonstrations Session, Prague, Czech Republic, 2007.

[10] K. Hassanzadeh, M. Reformat, W. Pedrycz, I. Jamal and J. Berezowski, "T2R: System for Converting Textual Documents into RDF Triples," in Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on , Atlanta, GA USA, 2013.

[11] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu and D. Jurafsky, “Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules,” Computational Linguistics, vol. 39, no. 4, pp. 885-916, 2013.

[12] S. Karmakar, Syntactic and Semantic Analysis and Visualization of Unstructured English Texts, Atlanta, Georgia : Doctor of Philosophy (PhD), Georgia State University, 2011.

[13] M. C. De Marneffe, B. MacCartney and C. D. Manning, "Generating Typed Dependency Parses from Phrase Structure Parses," in 5th International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, 2006.

[14] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, "Natural Language Processing (Almost) from Scratch," The Journal of Machine Learning Research, vol. 12, pp. 2493-2537, 2011.

[15] I. Augenstein, S. Pad´o and S. Rudolph, "LODifier: Generating Linked Data from Unstructured Text," in 9th Extended Semantic Web Conference (ESWC 2012), Greece, 2012.

[16] H. Brian and C. Stephen, "ASKNet: automated semantic knowledge network," in Proceedings of TwentySecond National Conference on Artificial Intelligence 889–894, Vancouver, BC , 2007.

[17] E. Peter and N. Pierre, Entity Extraction: From Unstructured Text to DBpedia RDF Triples, Boston, USA: The Web of Linked Entities Workshop, 2012.

[18] D. G. P. K. Martha Palmer, "The Proposition Bank: A Corpus Annotated with Semantic Roles," Computational Linguistics Journal, vol. 31, no. 1, pp. 71-106, 2005.

[19] C. Ramakrishnan, K. J. Kochut and A. P. Sheth, "A framework for schema-driven relationship discovery from unstructured text," in The Semantic Web-ISWC 2006. Springer Berlin Heidelberg 583-596, 2006.

[20] A. Gangemi, "A Comparison of Knowledge Extraction Tools for the Semantic Web," in The Semantic Web: Semantics and Big Data. Springer Berlin Heidelberg 351-366, 2013.

[21] A. Y. Mohamed, H. Johannes, B. Ilaria, S. Marc and W. Gerhard, " Aida: An online tool for accurate disambiguation of named entities in text and tables," in In Proceedings of the 37th International Conference on Very Large Databases, VLDB , Seattle, WA, US, 2011.

[22] P. Arambulo, "International programs and veterinary public health in the Americas--Success, challenges, and possibilities," Preventive veterinary medicine, vol. 86, no. 3-4, pp. 208-215, 2008.

[23] R. L. Berkelman, "Emerging infectious diseases in the United States, 1993," Journal of Infectious Diseases, vol. 170, no. 2, pp. 272-277, 1994.

[24] A. a. M. P. Ammon, “Integrated data collection on zoonoses in the European Union, from animals to humans, and the analyses of the data,” International journal of food microbiology, vol. 139, pp. s43-s47, 2010.

[25] J. S. Ahmed, H. A. Alp and U. M. and Seitzer, “Animal transboundary diseases: European Union and Asian network of veterinary research cooperation for quality livestock production,” Journal of Veterinary Medicine, vol. 53, no. Suppl. 1, pp. 2-6, 2006.

[26] I. AlchemyAPI, “Disambiguation,” AlchemyAPI, Inc., 2014. [Online]. Available: http://www.alchemyapi.com/api/entity/disamb.html. [Accessed 01 05 2014].

[27] S. C. a. J. R. Curran, “Parsing the WSJ using CCG and Log-Linear Models,” in 42nd Meeting of the ACL, pages 104–111, Barcelona, Spain., 2004.

[28] J. Boss, “Wide-Coverage Semantic Analysis with Boxer,” in Proceedings of the 2008 Conference on Semantics in Text Processing, PA, USA, 2008.

[29] O. d. b. G. W. |. A. b. J. Curran, “C&C tools - Trac,” c2006. [Online]. Available: http://svn.ask.it.usyd.edu.au/trac/candc/wiki.. [Accessed 01 05 2014].

[30] T. Shoemaker, A. MacNeil, S. Balinandi, S. Campbell, J. F. Wamala, L. K. McMullan, R. Downing, J. Lutwama, E. Mbidde, U. Stroher, P. E. Rollin, Nichol and S. T., “Reemerging Sudan Ebola virus disease in Uganda, 2011,” Emerging Infectious Diseases, vol. 18, no. 9, pp. 1480-1483, 2012.

[31] T. E. C. o. t. I. D. S. o. A. E. I. Network, “The Emerging Infections Network: a new venture for the Infectious Diseases Society of America,” Clinical Infectious Diseases, vol. 25, no. 1, pp. 34-36, 1997.

[32] K. M. De Cock, H. W. Jaffe and J. W. Curran, “Reflections on 30 years of AIDS,” Emerging Infectious Diseases, vol. 17, no. 6, pp. 1044-8, 2011.

[33] D. R. Harper, “Preparedness for SARS in the UK in 2003,” Philosophical Transactions of the Royal Society of London - Series B: Biological Sciences, vol. 350, no. 1447, pp. 1131-1132, 2004.

[34] K. H. Zessin, “Emerging Diseases: A Global and Biological Perspective,” Journal of veterinary medicine. B, vol. 53, no. s1, pp. 7-10, 2006.

[35] L. Cibrelus, V. Noel, J. Emmanuelli, G. Breton, P. Longuet, B. Rigolli, C. Leport and J. L. Vilde, “Management of 90 patients presenting with suspected severe acute respiratory syndrome. Experience of a collaboration between epidemiologists and clinicians facing an

emerging infectious disease health alert.," Medecine et Maladies Infectieuses, vol. 37, no. 3, pp. S242-S250, 2007.

[36] M. Recasens, M.-C. De Marneffe and C. Potts, "The Life and Death of Discourse Entities: Identifying Singleton Mentions," in In Proceedings of NAACL 2013 , Atlanta, 2013.

[37] N. Imp, "Exploring Stanford's CoreNLP Coreference system — Nectarine Imp," c2014. [Online]. Available: http://www.nectarineimp.com/2012/analytics/exploring-stanfords-corenlp-coreference-system/. [Accessed 01 06 2014].

[38] M. Cranfield and R. Minnis, "An integrated health approach to the conservation of Mountain gorillas Gorilla beringei beringei," International zoo yearbook, vol. 41, no. 1, pp. 110-121, 2007.

[39] I. OpenLink Software, "OpenLink virtuso universal server," c2014. [Online]. Available: http://virtuoso.openlinksw.com/. [Accessed 01 06 2014].

[40] "An Overview of Virtuoso Universal Server | DATAVERSITY," DATAVERSITY Education, LLC, 2013. [Online]. Available: http://www.dataversity.net/an-overview-of-virtuoso-universal-server/. [Accessed 01 06 2014].

[41] Gephi.org, "Gephi makes graphs handly," c2014. [Online]. Available: http://gephi.github.io/. [Accessed 01 06 2014].

[42] A. Ovens, "Virtuoso Importer - Gephi Marketplace," 30 05 2013. [Online]. Available: https://marketplace.gephi.org/plugin/virtuoso-importer/. [Accessed 01 05 2014].

[43] wikipedia, "simplicity definition," 2014. [Online]. Available: http://en.wikipedia.org/wiki/Simplicity. [Accessed 01 06 2014].

[44] J. Blancou, B. B. Chomel, A. Belotto and F. X. Meslin, "Emerging or re-emerging bacterial zoonoses: factors of emergence, surveillance and control," Veterinary Research, vol. 36, no. 3, pp. 507-522, 2005.

[45] D. B. Blossom and L. C. McDonald, "The challenges posed by reemerging Clostridium difficile infection," Clinical Infectious Diseases, vol. 45, no. 2, pp. 222-227, 2007.

[46] O. Restif, D. T. S. Hayman, J. R. C. Pulliam, R. K. Plowright, D. B. George, A. D. Luis, A. A. Cunningham, R. A. Bowen, A. R. Fooks, T. J. O'Shea, J. L. N. Wood and C. T. Webb, "Model-guided fieldwork: practical guidelines for multidisciplinary research on wildlife ecological and epidemiological dynamics," Ecology Letters, vol. 15, no. 10, pp. 1083-1094, 2012.

[47] K. Takatori and J. Kosuge, “Some remarks on mycotic zoonoses from veterinary medicine,” Japanese Journal of Medical Mycology, vol. 44, no. 4, pp. 249-251, 2003.

[48] S. Specter, L. Schuermann, C. Hakiruwizera and M. S. Sow, “ASM LabCap's contributions to disease surveillance and the International Health Regulations (2005),” BMC Public Health, vol. 10 Suppl 1, no. s7, 2010.

[49] J. G. Rodriguez-Torres, “International approach to eradication and surveillance for foot-and-mouth disease in the Americas,” Annals of the New York Academy of Sciences, vol. 916, pp. 194-198, 2000.

[50] K. L. Russell, J. Rubenstein, R. L. Burke, K. G. Vest, M. C. Johns, J. L. Sanchez, W. Meyer, M. M. Fukuda and D. L. Blazes, “The Global Emerging Infection Surveillance and Response System (GEIS), a U.S. government tool for improved global biosurveillance: a review of 2009,” BMC Public Health, vol. 11 Suppl 2, no. S2, 2011.

[51] D. O. Freedman, P. E. Kozarsky, L. H. Weld and M. S. Cetron, “GeoSentinel: the global emerging infections sentinel network of the International Society of Travel Medicine,” Journal of Travel Medicine, vol. 6, no. 2, pp. 94-98, 1999.

[52] S. Specter, L. Schuermann, C. Hakiruwizera and M. S. Sow, “ASM LabCap's contributions to disease surveillance and the International Health Regulations (2005),” BMC Public Health, vol. 10 Suppl 1, p. s7, 2010.

[53] J. W. Snyder, “The laboratory response network: before, during, and after the 2001 anthrax incident,” Clinical Microbiology Newsletter, vol. 27, no. 22, pp. 171-175, 2005.

[54] B. Monday, S. N. Gitta, P. Wasswa, O. Namusisi, A. Bingi, M. Musenero and D. Mukanga, “Paradigm shift: contribution of field epidemiology training in advancing the "One Health" approach to strengthen disease surveillance and outbreak investigations in Africa,” The Pan African medical journal, vol. 10 Supp 1, p. 13, 2011.

[55] M. C. Johns, R. L. Burke, K. G. Vest, M. Fukuda, J. A. Pavlin, S. K. Shrestha, D. C. Schnabel, S. Tobias, J. A. Tjaden, J. M. Montgomery, D. J. Faix, M. R. Duffy, M. J. Cooper, J. L. Sanchez, D. L. Blazes and A.-G. O. R. Group, “A growing global network's role in outbreak response: AFHSC-GEIS 2008-2009,” BMC Public Health, vol. 11 Suppl 2, p. S3, 2011.

Appendices

Appendix I: Senna Sample Output

The analyzed sentence: *"In the southwestern USA, hantavirus was recognized as the cause of a pulmonary syndrome with a mortality rate exceeding 50%."*

Senna output:

In	IN	S-PP	O	-	B-AM-LOC	O	(S1(S(PP*
the	DT	B-NP	O	-	I-AM-LOC	O	(NP*
southwestern	JJ	I-NP	O	-	I-AM-LOC	O	*
USA	NNP	E-NP	S-LOC	-	E-AM-LOC	O	*))
,	,	O	O	-	O	O	*
hantavirus	NN	S-NP	O	-	S-A1	O	(NP*)
was	VBD	B-VP	O	-	O	O	(VP*
recognized	VBN	E-VP	O	recognized	S-V	O	(VP*
as	IN	S-PP	O	-	B-A2	O	(PP*
the	DT	B-NP	O	-	I-A2	O	(NP(NP*
cause	NN	E-NP	O	-	I-A2	O	*)
of	IN	S-PP	O	-	I-A2	O	(PP*
a	DT	B-NP	O	-	I-A2	B-A0	(NP(NP*
pulmonary	JJ	I-NP	O	-	I-A2	I-A0	*
syndrome	NN	E-NP	O	-	E-A2	I-A0	*)
with	IN	S-PP	O	-	B-AM-MNR	I-A0	(PP*
a	DT	B-NP	O	-	I-AM-MNR	I-A0	(NP(NP*
mortality	NN	I-NP	O	-	I-AM-MNR	I-A0	*
rate	NN	E-NP	O	-	I-AM-MNR	E-A0	*)
exceeding	VBG	S-VP	O	exceeding	I-AM-MNR	S-V	(VP*
50%.	.	S-ADJP	O	-	E-AM-MNR	S-A1	(NP*))))))))))))

Appendix II: Alphabetical list of part-of-speech tags used in the Penn Treebank Project

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential *there*
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	*to*
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Appendix III: The C&C output:

The analyzed sentence “Another visible accomplishment is the elimination of hydatidosis in the endemic countries and regions of the southern cone . “

```
% this file was generated by the following command(s):
%   bin/candc --models models/boxer

:- multifile w/8, ccg/2, id/2.
:- discontiguous w/8, ccg/2, id/2.
:- dynamic w/8, ccg/2, id/2.

ccg(1,
 rp('S[dcl]',
  ba('S[dcl]',
   fa('NP[nb]',
    lf(1,1,'NP[nb]/N'),
    fa('N',
     lf(1,2,'N/N'),
     lf(1,3,'N'))),
   fa('S[dcl]\NP',
    lf(1,4,'(S[dcl]\NP)/NP'),
    ba('NP',
     fa('NP[nb]',
      lf(1,5,'NP[nb]/N'),
      lf(1,6,'N')),
     fa('NP\NP',
      lf(1,7,'(NP\NP)/NP'),
      ba('NP',
       lex('N','NP',
        lf(1,8,'N')),
       fa('NP\NP',
        lf(1,9,'(NP\NP)/NP'),
        ba('NP',
         fa('NP[nb]',
          lf(1,10,'NP[nb]/N'),
          fa('N',
           lf(1,11,'N/N'),
           ba('N',
            lf(1,12,'N'),
            conj('conj','N','N\N',
```

```
            lf(1,13,'conj'),
            lf(1,14,'N'))))),
          fa('NP\NP',
           lf(1,15,'(NP\NP)/NP'),
           fa('NP[nb]',
            lf(1,16,'NP[nb]/N'),
            fa('N',
             lf(1,17,'N/N'),
             lf(1,18,'N'))))))))))),
  lf(1,19,'.'))).

w(1, 1, 'Another', 'another', 'DT', 'I-NP', 'O', 'NP[nb]/N').
w(1, 2, 'visible', 'visible', 'JJ', 'I-NP', 'O', 'N/N').
w(1, 3, 'accomplishment', 'accomplishment', 'NN', 'I-NP', 'O', 'N').
w(1, 4, 'is', 'be', 'VBZ', 'I-VP', 'O', '(S[dcl]\NP)/NP').
w(1, 5, 'the', 'the', 'DT', 'I-NP', 'O', 'NP[nb]/N').
w(1, 6, 'elimination', 'elimination', 'NN', 'I-NP', 'O', 'N').
w(1, 7, 'of', 'of', 'IN', 'I-PP', 'O', '(NP\NP)/NP').
w(1, 8, 'hydatidosis', 'hydatidosis', 'NN', 'I-NP', 'O', 'N').
w(1, 9, 'in', 'in', 'IN', 'I-PP', 'O', '(NP\NP)/NP').
w(1, 10, 'the', 'the', 'DT', 'I-NP', 'O', 'NP[nb]/N').
w(1, 11, 'endemic', 'endemic', 'JJ', 'I-NP', 'O', 'N/N').
w(1, 12, 'countries', 'country', 'NNS', 'I-NP', 'O', 'N').
w(1, 13, 'and', 'and', 'CC', 'O', 'O', 'conj').
w(1, 14, 'regions', 'region', 'NNS', 'I-NP', 'O', 'N').
w(1, 15, 'of', 'of', 'IN', 'I-PP', 'O', '(NP\NP)/NP').
w(1, 16, 'the', 'the', 'DT', 'I-NP', 'O', 'NP[nb]/N').
w(1, 17, 'southern', 'southern', 'JJ', 'I-NP', 'O', 'N/N').
w(1, 18, 'cone', 'cone', 'NN', 'I-NP', 'O', 'N').
w(1, 19, '.', '.', '.', 'O', 'O', '.').
```