*machine learning & knowledge extraction*

MDPI

*Article*

# Knowledge Graphs Representation for Event-Related E-News Articles

**M.V.P.T. Lakshika * and H.A. Caldera**

University of Colombo School of Computing (UCSC), University of Colombo, Colombo 00700, Sri Lanka; hac@ucsc.cmb.ac.lk

* Correspondence: lakshikamvpt@gmail.com

**Abstract:** E-newspaper readers are overloaded with massive texts on e-news articles, and they usually mislead the reader who reads and understands information. Thus, there is an urgent need for a technology that can automatically represent the gist of these e-news articles more quickly. Currently, popular machine learning approaches have greatly improved presentation accuracy compared to traditional methods, but they cannot be accommodated with the contextual information to acquire higher-level abstraction. Recent research efforts in knowledge representation using graph approaches are neither user-driven nor flexible to deviations in the data. Thus, there is a striking concentration on constructing knowledge graphs by combining the background information related to the subjects in text documents. We propose an enhanced representation of a scalable knowledge graph by automatically extracting the information from the corpus of e-news articles and determine whether a knowledge graph can be used as an efficient application in analyzing and generating knowledge representation from the extracted e-news corpus. This knowledge graph consists of a knowledge base built using triples that automatically produce knowledge representation from e-news articles. Inclusively, it has been observed that the proposed knowledge graph generates a comprehensive and precise knowledge representation for the corpus of e-news articles.

**Keywords:** knowledge graph; knowledge base; knowledge representation; e-news articles; SPO triples



**Citation:** Lakshika, M.V.P.T.; Caldera, H.A. Knowledge Graphs Representation for Event-Related E-News Articles. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 802–818. https://doi.org/10.3390/make3040040

Academic Editor: Andreas Holzinger

Received: 15 August 2021
Accepted: 20 September 2021
Published: 26 September 2021



## 1. Introduction

Modern technologies allow e-newspapers to offer their readers more interactive and inexplicable news content that was not possible in the printed medium. Thereby, printed newspapers are slowly losing their gleam, while e-newspapers continue to grow with the availability of quick and obsolete news to the world with only a few snaps. With the impact of e-newspapers on modern life, the introduction of e-news websites has facilitated many web news content creators to generate news information. E-news articles in such e-news websites include hundreds of semi-structured and unstructured text contents. The internet is continuously flooded with many news websites and web news articles of various lengths of size that often mislead people who read and understand them. Additionally, the quantity of web news information available on a particular news event is enormous. The worst situation faced by the e-news readers is when multiple e-news articles report the same news event with different levels of information content. Most of the time, e-newspaper readers are facing situations where they need to form opinions, shape beliefs, or take decisions on certain news events in these e-news articles. Hence, this has forced e-newspaper readers to only obtain important information from these large amounts of web news information, which is a very procrastinate and tiresome task. As a solution for this issue, this paper presents an approach that automatically extracts only the essence from e-newspapers that include a huge amount of obsolete news.

Information representation techniques that can speedily represent only the important information in a collection of e-news are required to address the above issue. Developing computational models for automatic summarization is one of the fundamental types of

information representation that can represent the gist of a text document, and it has considerably attracted the attention of many researchers. The current automatic text summarization methods are based on traditional approaches, most of which extract only the keywords or key sentences in a text document [1], giving more importance to the words that are more frequent in the text document. Even though the general sequence-to-sequence text extraction methods [1] have significant improvement compared to the traditional methods, they cannot be easily accommodated with the contextual information of a news event, such as background information that provides a broader understanding of an event, person, or item in a news document to obtain higher-level abstraction [1]. News events in e-news articles that are related to the equivalent entities may extend over multiple sentences and multiple articles, making them challenging for existing sequential models to address this complexity [2]. To capture useful insights from a collection of e-news articles that include texts with a similar news event, there should be an efficient and well-organized approach to represent all these textual data contents.

Following the existing research gaps in generating highly abstractive information representations, including contextual information of a news event, the authors conducted a study to develop a knowledge-based text representation approach that builds a knowledge graph (KG) for knowledge representation in a set of e-news articles based on a user query. Knowledge Graphs [1–13] are an efficient and powerful knowledge representation technique in the research community. The generation of automatic summaries using Knowledge-Graph-based approaches, which are capable of capturing human experts' knowledge to support the decision-making process, is still hardly touched due to their complexity.

While many pieces of research about building a general-purpose and domain-specific KG have been published, we focus on building a scalable KG that represents the most commonly known knowledge in a collection of e-news articles based on e-news events. This paper investigates whether a KG can be utilized as an efficient application in analyzing and generating knowledge representation using a set of e-news articles based around one news event and construct and employing a scalable KG as an approach in realizing the solution that automatically extracts only the essence from a huge amount of obsolete news in e-newspapers. The underlying process that we propose consists of three high-level steps: (1) extraction of the news content from e-news articles, (2) identification of the basic elements in KG and modeling the KG, and (3) finally, building the KG to visualize knowledge representation.

This study makes the following contributions:

- Introduction of a KG pipeline, which builds a KG for knowledge representation in a set of e-news articles based on a user query.
- Authors in this study guarantee the comprehensiveness and preciseness of knowledge representation by inferencing missing relational links and new relational links in e-news articles using statistical relational learning (SRL).

The KG-based text representation approach presented in this paper would be beneficial for the e-newspaper readers to speedily comprehend the dominant idea of a set of e-news article collections related to the same news event. Thus, our approach meets the urgent need for a knowledge-based text representation approach that can automatically extract and represent information from a collection of e-news articles.

The rest of this paper is ordered as follows: Section 2 presents the introduction to the related works of widely known and published KGs. Section 3 presents the general architecture of modeling KGs. Section 4 elaborates on the proposed solution to construct a KG pipeline, and Section 5 presents the performance evaluation of the resulting KG based on a user query. Discussions are elaborated in Section 6, and finally, all the conclusions, limitations, and future directions are described in Section 7.

## 2. Literature Review

Over 80% of data in the World Wide Web (WWW) are present in the form of an unstructured or semi-structured type, which makes knowledge extraction one of the major challenges in the KG research community. Fifteen years ago, researchers at the University of Twente and the University of Groningen, Netherlands, initiated the researches on KG theory and continued its focus on KG applications to analyze a text [5]. Over the years, several open KBs or ontologies have been created, including WordNet, DBpedia, YAGO, Freebase, BabelNet, and NELL, covering millions of real-world entities and relations in different domains [3,10,11,13]. Among these, traditional KGs such as ConceptNet and Freebase usually contain known facts and assertions about entities [12]. Freebase was released as an openly shared database in 2007, and its data were collected from MusicBrainz, Wikipedia, Fashion Model Directory and were also contributed by its general users. Later, all the data in Freebase, which consists of more than 2 billion triples, 7000 properties, and 4000 types, were transferred to Wikidata [11]. In 2012, Wikidata was firstly published, and it has been publicly maintained till now with a collection of more than 55 million entities derived from Wikivoyage, Wikipedia, Wikisource, and other websites. DBpedia is one more large-scale KG based on Wikipedia and also collaboratively modified by the general community. It is a multilingual KG that is available in 125 languages and has more than 40 million entities. ConceptNet, launched in 1999, is another open multilingual KG that contains more than 8 million entities with nearly 21 million entity links. YAGO is another open-sourced semantic KG resultant from Wikipedia and WordNet, which has more than 10 million entities and 120 million entity links. The concept of KG expanded its popularity in 2012 with the first KG known as the 'Knowledge Vault', launched by Google's search engine, which is a knowledge fusion framework that was proposed to build large-scale KGs [10]. The potential data requirement of such well-constructed largescale KGs is much larger but can be advantageous for many applications.

Most of the data amongst millions of entities and triples in large-scale KGs, including Google's search engine, are extracted from Wikipedia data sources. Even with the existence of such sturdy data collection, those KGs could not sustain the requirements in question answering, text summarization, recommendation systems, and other web search applications. For instance, the latest COVID-19 vaccination information in Sri Lanka, which can only be obtained from some specific websites on the internet, is not held in the existing large-scale KGs. Apart from that, none of the existing KGs discuss the scalability of KGs to compact and extend with event-specific text data as in news articles.

Usually, KGs are constructed using predefined databases or ontologies built by domain experts, human-annotated training data or semi-structured textual sources, or unstructured text on the web using manual curation approaches and automated techniques. Manual curation requires domain experts to study text sources and annotate sentences that include relationships, which is a tiresome and time-consuming process [3,13]. Manual approaches limit the scalability and applicability of KGs across various domains, while automated approaches which depend on machine learning or NLP techniques rapidly discover sentences of interest. The Open Information Extraction (OpenIE) technique, which extracts all semantic relational tuples in source text without human supervision, has been introduced to overcome the limitations in manual curation [3]. Later, researchers attempted to address the limitations in manually curated KGs by developing a procedure to populate a KB semi-automatically or automatically with values extracted from a corpus. These automated approaches can be categorized as unsupervised and supervised machine learning approaches and rule-based extractions which will be discussed later in this paper [13].

Wang et al. [11] presented a KG developed by the Sogou Web Search Department, as none of the above-mentioned large-scale KGs completely fulfills the requirements of Sogou Inc. The Sogou KG is built with three applications of KG in Sogou Inc., Beijing, China: (1) entity detection and linking, (2) question answering, and (3) a knowledge-based dialog system to support efficient online products sales. The Sogou KG is constantly being updated, and currently, it comprises more than 54 million entities and 600 million

entity links extracted from 136 different websites. Lately, there has been growing interest in generating domain-specific knowledge representations, and to address the subject's specificity in these KGs, they have often been constructed by integrating databases that were populated by domain experts via manual curation techniques [13]. Crunchbase is one of such manually curated financial KBs covering 100,000+ business entities and business transactions. To address the limitations in manually curated domain-specific knowledge representations, automatic knowledge extractions and knowledge-representation methods have been introduced. Recent research by Elhammadi et al. [3] designed a high-precision automatic knowledge extraction pipeline to build a KB semi-automatically or automatically with quintuples of the form subject–predicate–object–monetary value–date extracted from a financial news domain using Semantic Role Labeling (SRL) and pattern-based information extraction to extract noun- and verb-mediated relations. This pipeline contains over 380,000 extracted triples, which are readily queried.

Nicholson et al. [13] presented a more robust automated KG construction approach based on machine learning techniques to build a low-dimensional representation that supports biomedical applications. This representation is designed to preserve both the local and global structure of KG while using machine learning methods to generate predictions within genomic, clinical, and pharmaceutical domains. Even though several techniques are used to build KGs, often, machine learning techniques are used to construct a low-dimensional representation that still supports many different applications [13]. The abstractive summarization with graph augmentation and semantic-driven reward (ASGARD) framework developed by Huang et al. [2] used a regular document encoder along with a separate graph-structured encoder to preserve both global context and local characteristics of entities in text documents using the outputs taken from an open information extraction (OpenIE) system.

Motivated by the inconsistency between the reader's limited information desires and the enormous scale of KGs, a novel solution called personalized KG summarization [4,8], has been raised. The increasing scale of all-encompassing KGs emphasized the necessity of summarization as a method of efficiently accessing and distilling world knowledge. A personalized KG constructs a compressed personal summary from a KG that contains only the relevant facts of the reader's interests and queries [4]. Such personalized KGs support fast, interactive querying, and they are constantly being updated to serve to change user needs and data evolution. A personalized KG is a summary or sparse graph that maximizes a user's inferred 'utility' over a given KG with regard to a user and device-specific constraint on the summary's size. Another personalized KG approach by Safavi et al. [8] is GLIMPSE, which is a summarization framework that offers theoretical guarantees on the summary's usefulness. A KG constructed by Wu et al. [1] can not only obtain higher-level extraction from the text but also can select a template and question and answer to obtain a personalized abstractive summary from the source documents [1].

Generating a text summary from a collection of text documents retrieved based on the query given by a user [6,7] using semantic knowledge extraction techniques is another application in KG. The Summarized Research Article Generator (SRAG) introduced by Malviya et al. [6] stores the semantic knowledge of each research paper, which is extracted through a query in the form of a semantic tree. A probabilistic model based on a Bayesian network is used to extract the relevant information from the semantic tree, and finally, the summarized article is generated segment-wise by merging only the most relevant paragraphs to maintain coherency.

Al-Khatib et al. [12] study the end-to-end construction of an argumentation KG that is intended to support argumentative question and answering, argument synthesis, and fake news detection with proven effectiveness for interpretable and controllable text generation and exploratory search. This framework automatically identifies all modeled types of instances based on a corpus that comprises about 16,000 manual annotations of 4740 claims with instances of the model's elements. To construct a KG from an input document, Huang et al. [2] utilize Stanford CoreNLP to obtain outputs from co-reference resolution

and OpenIE models without incorporating global entity linking across documents. They further collapse co-referential mentions of the same entity into one node for localizing salient contents related to each entity in the documents and thereby identify associates of spread-out entities through the graph paths [2].

## 3. Modeling Knowledge Graph

### *3.1. Knowledge Graph*

Knowledge Graph (KG) is one of the novel methods of Natural Language Processing (NLP) for Knowledge Representation (KR), which is very important for the problems that require taking a series of conclusions from existing text [5]. A KG is a large network of interconnected data constructed through Knowledge Bases (KB), which gather information from source text on web pages, databases, audio, and video contents. KG demonstrates the human language by analyzing entities and their relationships in KB factually and also augmented with background knowledge of an event [7]. KBs can either be domain-specific or generic [5], and in practice, a KB includes loads of triples, which are also termed as facts. A generic KB has a broader coverage of general worldly facts which do not oblige their knowledge to a particular domain.

KG is an instrument to represent the conceptual structure of a text document representing an idea, event, situation, or circumstances, which comprises nodes to represent concepts and edges to represent the conceptual relationships among the nodes. Nodes and edges in the KG are the main elements to represent the logical structure of knowledge expressed in natural language. Most research efforts have given a definition for KG by describing essential characteristics or general semantic representations in text. Nevertheless, there is no such proper definition for a KG. Following preceding literature, we outline a Knowledge Graph, G = {E, R, F} where E, R, and F are sets of entities, relations, and facts, respectively. A fact, f $\epsilon$ F is denoted as a triple (h, r, t) where h, r and t are head, relation, and tail, respectively. The nodes and edges in the KG are represented in the form of Subject–Predicate–Object (SPO) triples, where subject and object are entities, while the predicate is the relationship amongst those entities [3]. This collection of nodes and edges can be appropriately stored and maintained as a KG. Modern KGs comprise up to billions of entities and relationships and also are repetitively being enlarged with the occurrence of novel facts. Moreover, these KGs will probably only cover a portion of individual readers' information desires. Due to that, the necessity of summarization is heightened, which reduces the knowledge in these huge KGs to a more concise representation but to the point that they are still interpretable.

In linguistics, semantics is the subfield that studies the meaning of words, phrases, and sentences. Semantics are closer than logic in a sentence, and semantics are a part of the formal language that highlights the relationship between form and meaning within a sentence [5]. Hence, KGs can be used to describe human insight in the semantic model, which contains general knowledge about the world and process information. KG represents the knowledge in the source documents and can be considered as an overview that defines and explains the natural language more than the aspect of semantics for syntax [5]. KG representation identifies the most significant information present in the source documents and represents that information to the reader in a condensed manner based on the individual readers' information desires. KG theory is important in the formation of knowledge out of information in a text document. Therefore, the research community continues studies to analyze and develop KG representations to recognize important information from a collection of texts sources that can be utilized as required knowledge in the context of decision making. There is great concentration in extracting high-quality knowledge in applications such as question answering, information searching, chatbots, web applications, knowledge management systems, biomedical applications, social network mining, and recommender systems [3,13,14].

The generation of concise knowledge representation from e-news articles requires a well-structured representation of source text to identify the interaction of relevant enti-

ties [2] in news articles. Since knowledge representation in a KG generates a structured summary while highlighting the proximity of related concepts in a collection of text documents [2], KG can be used as an efficient application in representing knowledge in a set of e-news articles based around a particular news event. One of the most important linguistic relationships that can be found within a KG is the hierarchical relationship "hypernyms" [15,16]. For example, the "is–a" relation of conceptual ontologies relates subordinate words (e.g., "Sri Lanka") with its superordinate (e.g., "Country"). Hierarchical relation lies at the origin of human reasoning and permits relating words representing specific instances to generic ones; moreover, it is one of extreme importance in any taxonomy [16].

### *3.2. Knowledge Graph Construction*

Data extraction is the first step of automatically constructing a KG by extracting data into a predefined form from an unstructured or semi-structured text in text documents or crawled from the internet such as web pages, JSON data downloaded by APIs, or XML data. Wang et al. [11] store the output data in the form of JSON-LD that records the data extracted from the input data. Knowledge representation in KG joins many knowledge-acquisitions tasks, which are the step of constructing KG from text sources or completing an existing KG by discovering and recognizing more entities and relations [10]. Relation extraction, entity recognition, and entity alignments—which can be considered entity-oriented acquisition tasks—and Knowledge Graph Completion (KGC) [10] are the main tasks in the knowledge acquisition process.

#### 3.2.1. Relationship Extraction

Relation extraction is a key task of building large-scale KGs automatically by extracting unidentified relations from the source text. Relationships can be considered as unidirectional (e.g., the Sinopharm vaccine prevents COVID-19, but COVID-19 cannot prevent Sinopharm); however, there are circumstances where relationships can be considered as bidirectional (e.g., a virus resembles another virus) [13]. Automated relationship extraction approaches can be categorized as supervised machine learning, unsupervised machine learning, and rule-based extraction approaches [13].

Rule-based approaches, which involve recognizing important grammatical patterns and keywords in the sentences to identify relationships between words, are the foundation for many relationships extraction systems. Keywords are identified with the support of knowledge experts or using predefined ontologies, while grammatical patterns are identified using parse trees curated by domain experts [13]. Both approaches require a substantial amount of manual exertion and expert domain knowledge to perform well. The future direction in rule-based approaches is to automate the manual construction of these hand-crafted patterns and accelerate the process of rule-based relationships extraction systems [13]. Parser trees are used by many text mining and relationships extraction approaches to generate features for machine learning algorithms that use rule-based extraction to detect sentences with relationships. Unsupervised relationship extraction approaches identify inferences from textual data without using annotated labels. These methods comprise some form of statistical calculations or clustering [13] to automatically extract relations from a corpus. Supervised relationship extraction approaches use labeled sentences to construct general patterns that detect sentences with positive relationships from sentences that do not have a negative relationship. Most of the approaches that use pre-labeled publicly available datasets succeeded and range from linear classifiers such as regression analysis to non-linear classifiers such as Support Vector Machines (SVMs) and deep learning methods.

#### 3.2.2. Entity-Oriented Acquisition

Entity discovery is an entity-based knowledge acquisition method that explores entity-related knowledge under different approaches such as entity recognition, entity linking, entity disambiguation, and entity typing.

Entity Recognition (ER) focuses on identifying specifically mentioned named entities in text documents and tags such entities. Capitalization patterns and language-specific resources such as gazetteers are several hand-crafted features applied in many pieces of literature. Entity disambiguation is another significant combined task which links ambiguous entity mentions to their referent entities in a KG. For example, "Sir Arthur Charles Clarke is an English writer, notable for both his science fiction and nonfiction". The entity mentions of "Arthur Charles Clarke" and "his" should be linked to the entity of "Arthur C. Clarke". The entity-linking task detects the relevant character string in natural language text that represents the entity and maps it to a specific entity in the KB [11]. This task supports identifying the correct entity from multiple candidates while eliminating ambiguity. For example, "Arthur C. Clarke" has multiple possible candidate entities, which may represent an English science-fiction writer, futurist, science writer, undersea explorer, inventor, or even a movie or book with an identical name. Accurately linking entities in such situations become more complex if the context information is poor. A well-designed entity-linking method considers many factors, including the preceding knowledge of the entity itself, the fit degree between the context in which the entity and the phrase are located, and the matching degree between the entity and the phrase [11]. Entity typing is another entity discovery approach that aims to identify the semantic type of an entity in the source text using coarse and fine-grained types, where the latter uses a tree structure.

#### 3.2.3. Knowledge Graph Completion (KGC)

KGC is the process of completing missing links between existing entities in the KG using subtasks such as entity prediction, link prediction, and relation prediction. Primary research on KGC focused on learning low-dimensional embedding for subject–verb–object triple prediction. Triple-prediction-based approaches mostly depend on triple representation learning to capture semantics and use candidate-ranking methods for completion. Reasoning in such approaches remains good at the individual relation level and is poor at complex reasoning due to the ignorance of the symbolical nature of KG. Most KGC approaches are unsuccessful in capturing multi-step relationships. Thus, recent research exertion turns to discover multi-step relation paths incorporating logical rules, rule-based reasoning, and termed relation path inference [10].

Facts or information extraction from source text frequently end up with erroneous facts. As a solution for this, Cohen et al. proposed a methodology to jointly evaluate the extracted facts [6], but the issue with this methodology was its consideration of only the trivial set of possible errors that could arise in extracted facts. Elhammadi et al. [3] proposed a Conditional Random Field (CRF) model to recognize and remove sequences of noisy text frequently found in financial news articles. Since their [3] KG is built for financial news documents, they build a dictionary of word-sense-disambiguated financial predicates to filter out noisy extractions made by Statistical Relational Learning (SRL) as a further process in improving precision.

### *3.3. Knowledge Graph Evaluation*

Evaluation in KGs remains difficult since the experiments in this context comprise a significant amount of time and resources such as domain experts [10,13]. Moreover, the multiplicity of KG applications obstructs the development of a standardized set of evaluations. The various biases related to the degree of nodes and their distributions must be considered for accurate evaluation in a KG [13]. Wang et al. [17] analyzed the knowledge representation based on the scoring functions while keeping a specific focus on the diverse type of information used in each knowledge representation. Recent research by Ji et al. [10] categorized and provided a full-scale view of knowledge representation considering four-folds; representation space, scoring function, encoding models, and auxiliary information. The scoring functions were generally used to measure the plausibility of factual triples in the KG representations, while Ji et al. [10] implemented the scoring function for the unification process. There are two distinctive types of scoring functions; distance-based and similarity-

based functions, where both are used to measure the plausibility of a fact in a KG. The most widely used scoring functions, which are based on distance, measure the plausibility of facts by calculating the distance between entities and relations as h + r ≈ t where h = head entity, r = relation, and t = tail entity in a triple [10]. Generally, semantic similarity-based scoring functions measure the plausibility of facts using semantic matching, while most widely used similarity-based scoring functions frequently adopt a multiplicative formulation, $h^T M_r \approx t^T$, where the mapping matrix $M_r \in \mathbb{R}^{k \times d}$ is a projection matrix from the entity space ($h,t \in \mathbb{R}^k$) into the relation space ($r \in \mathbb{R}^d$) to transform h near t in the representation space, and $\mathbb{R}$ is the relation matrix [10].

## 4. Research Methodology

The basic structure of the pipeline used in the KG construction process using e-news corpus is shown in Figure 1.

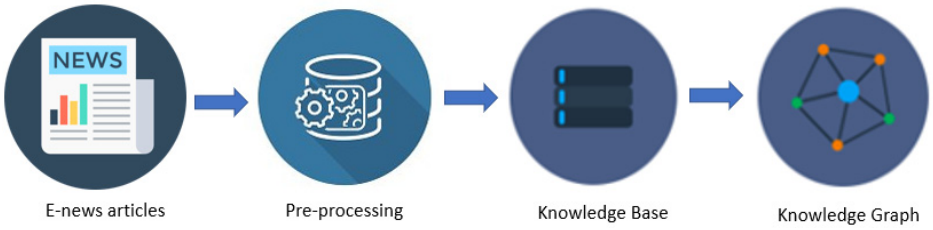


**Figure 1.** The KG construction pipeline.

### *4.1. Phase 1—E-News Article Extraction*

Source texts relevant for KG construction were identified from the unstructured and semi-structured event-related e-news articles in the e-news websites. Source text was crawled from web pages in several e-news websites such as Daily News, Daily Lankadeepa, Ceylon Today, Lanka News Web (LNW), and Hindustan times based on the user query. For example, "COVID-19 vaccination in Sri Lanka".

Python-goose, an article content extractor that extracts most likely article content from a URL or previously fetched HTML documents, was used to identify the source text content from the e-news corpus. While the source text extracted from e-news websites encompasses mostly free text, many data comprise structured information such as images, links to access external web pages, and geo-coordinates. Disambiguation pages were eliminated during the article content extraction process using Python-goose.

### *4.2. Phase 2—Preprocessing*

These source texts conclude the pre-processing of information for KB after identifying information or facts which contain the important parts of the original sentence. For example, one sentence taken from the resulting source text generated for the above user query is "Sri Lanka began COVID-19 vaccination for its frontline workers."

Initially, we clean the source text to filter necessary information by recognizing the entities, relationships associated with entities, and attributes, which are the properties that further describe entities and relations. This identification process utilizes several NLP techniques such as Named Entity Recognition (NER), Part of Speech (POS) tagging, lemmatization, stemming, co-reference resolution, and entity disambiguation. Results of applying NLP techniques are described in Figures 2 and 3.

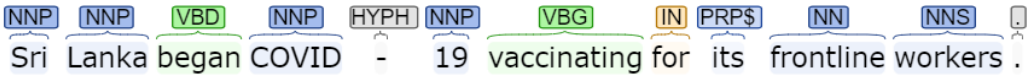


**Figure 2.** Tags generated by NLP technique, Part of Speech (POS) tagging.

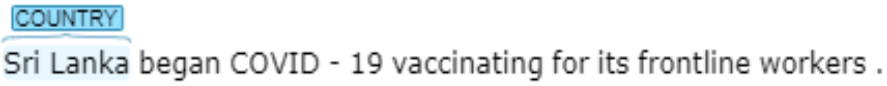


**Figure 3.** Identified entities after applying Named Entity Recognition (NER).

### *4.3. Phase 3—Knowledge Base Construction*

All the sentences in the source text are generalized in the form of triples and stored within the Knowledge Base. A triple is a composition of the subject, predicate, and its object (SPO) in a sentence, where both subject and object are entities that are involved in a relationship defined by the predicate. We utilized the Stanford CoreNLP libraries to obtain outputs from open information extraction (OpenIE) models as described in Figure 4 and populated the KB with the formulated triples.

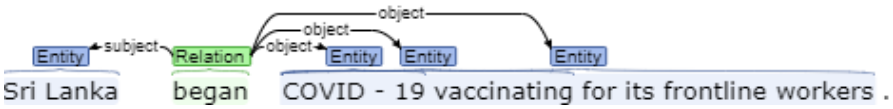


**Figure 4.** Obtained outputs from open information extraction (OpenIE) using the Stanford CoreNLP libraries.

The breakdown of triple for our example is as below.

Subject: Sri Lanka.

Predicate: began.

Object(s): {COVID-19}, {COVID-19 vaccinating}, {COVID-19 vaccinating for its frontline workers}.

We extract SPO triples using the rule-based extraction approach and eliminate any triple whose subject or object consists of more than 10 words. In occurrences such as two triples differ only by one argument and the arguments overlap, we possess only the longer triple inside the KB. So, within the KB, we will have the above statement in the form of began (Sri Lanka, COVID-19 vaccinating). This is a single triple within the KB.

Applying co-reference resolution, which is the process of discovering and replacing all expressions or phrases that denote the same entity in a source text with a common co-reference, is crucial as it is a significant fact for many higher-level NLP approaches that involve Natural Language Understanding (NLU) such as information extraction. We utilized the Stanford CoreNLP libraries to obtain outputs from co-reference resolution models, as described in Figure 5.

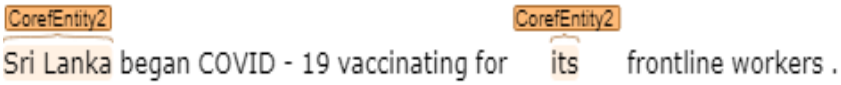


**Figure 5.** Obtained output from co-reference resolution using the Stanford CoreNLP libraries.

Relation Extraction using parse trees are commonly used to illustrate a sentence's grammatical construction using a tree data structure, and parse trees are presented in two arrangements: a constituency parse tree, which breaks a sentence into sub phrases (Figure 6), and a dependency parse tree, which analyzes the grammatical construction of a sentence (Figure 7).

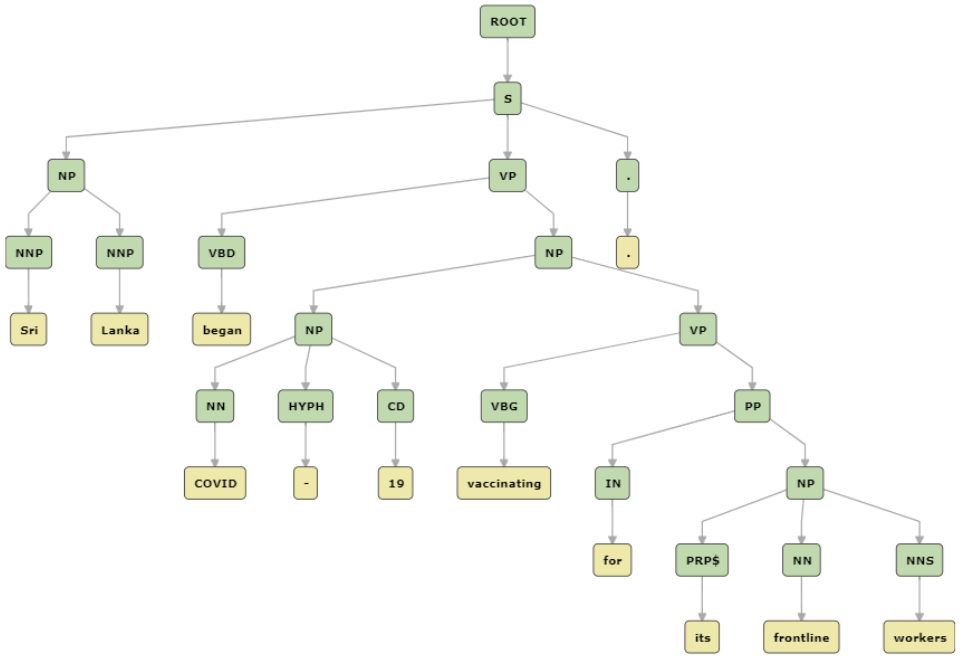


**Figure 6.** Constituency parse tree, which breaks a sentence into sub -phrases.

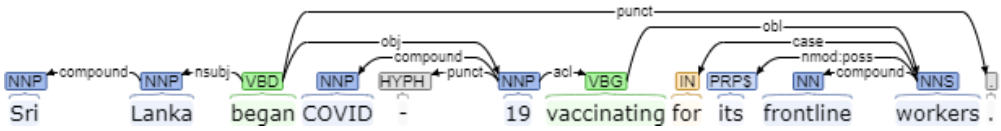


**Figure 7.** Dependency parse tree, which analyzes the grammatical construction of a sentence.

Both trees use POS tags and labels that dictate the grammatical role of a word, such as verb, noun, adverbs, adjective, conjunction, determiners, prepositions, etc., for construction of tree but represent the same information content in two different forms.

Hearst-pattern-based methods described in Table 1 have been one of the most influential approaches for recognizing hierarchical level relationships in text corpora using a set of predefined manually curated patterns to identify high-precision extractions [15]. For instance, a pattern such as "NPy such as NPx" or "NPx and other NPy" often indicates hypernym relations in the form of "x is-a y". Such patterns may be predefined, or they may be learned automatically. For instance, the sentence "Countries such as Sri Lanka received free vaccines." has the hypernym detection breakdown as below.

**Table 1.** Hearst patterns used in this study.

| Pattern |
|---|
| X which is a (class ∣ example ∣ kind ∣ . . . ) of Y |
| X (any ∣ some) (and ∣ or) other Y |
| X which is called Y |
| X is JJS (most)? Y |
| X a special case of Y |
| X is an Y that |
| X is a! (part ∣ member ∣ given) Y |
| ! (properties ∣ features) Y such as X1, X2, . . . |
| (Unlike ∣ like) (most ∣ any ∣ all ∣ other) Y, X |
| Y including X1, X2, . . . |

Relation: such as.
Hypernym: Countries.

Hyponym: Sri Lanka.

However, a well-recognized limitation in Hearst patterns is their sparsity: words related to the patterns must co-occur in a sentence exactly with the correct configuration; otherwise, no relation can be identified [15]. Since KG constructions that could incorporate all entity types are time-consuming [11], we emphasize only a limited type of entities that are important in the news domain, including DATE, LOCATION, MONEY, ORGANIZATION, PERCENT, PERSON, TIME.

Before moving to phase 4 of the pipeline, we inferred the missing links within the KB using SRL frameworks. SRL will be applied on the triples in KB to calculate confidence values for each triple as opposed to the entire domain in KB in order to identify how far those triples would hold accurate information. Missing links above the 0.8 confidence will be identified, and the newly inferred relational links will be formed and added to KB.

### *4.4. Phase 4—Knowledge Graph Construction*

Construction of the KG begins by considering subjects and objects in each triple as nodes connected by directed edges with predicates as attributes. Next, all the identified co-referential mentions of the same entity are combined into one node, and thereby, prominent contents related to each entity in the KG are localized, and it is possible to make interconnections among spread-out entities through graph paths [2].

A sample text crawled for the user query "COVID-19 vaccination in Sri Lanka" is shown below.

**Sri Lanka to start coronavirus vaccinations for general public from March**

PUBLISHED ON FEB 08, 2021 05:16 PM IST

Sri Lanka began vaccinating its frontline workers late last month after receiving 50,000 free Oxford-AstraZeneca-Covishield vaccines from India. Sri Lankan State Minister of Primary Health Care Sudarshini Fernandopulle on Monday said the coronavirus vaccination drive for the general public will begin in the first week of March, starting with people above the age of 60 years. Fernandopulle said the general public will be inoculated against the disease from next month. She said the drive will start with people over the age of 60 years and people in the workforce aged between 30 and 60 years. As of now, 1,60,000 frontline workers, including police and military personnel, have been vaccinated so far, Fernandopulle said. Apart from the vaccine gift from India, a consignment of vaccines would arrive in the island in early March through the COVAX international initiative. Sri Lanka will also import another 3 million vaccine doses from India.
As part of the immunization drive for the public, 4000 vaccination centers would be set up throughout the country with 2000 being operated daily where 300 people could be vaccinated every day.

Source: https://www.hindustantimes.com/
Accessed on 20 March 2021

Once new or missing relational links based on confidences are identified, we can build a KG with a set of highly confident facts. The resulting KG encompasses newly discovered facts based on the available observations that cannot be explicitly driven in addition to the original facts that were extracted during phase 3.

All the extracted triples and inferred relational links are stored in the KB and readily available for construction of the KG. A portion of sample KG generated by our pipeline for the above source is shown in Figure 8.

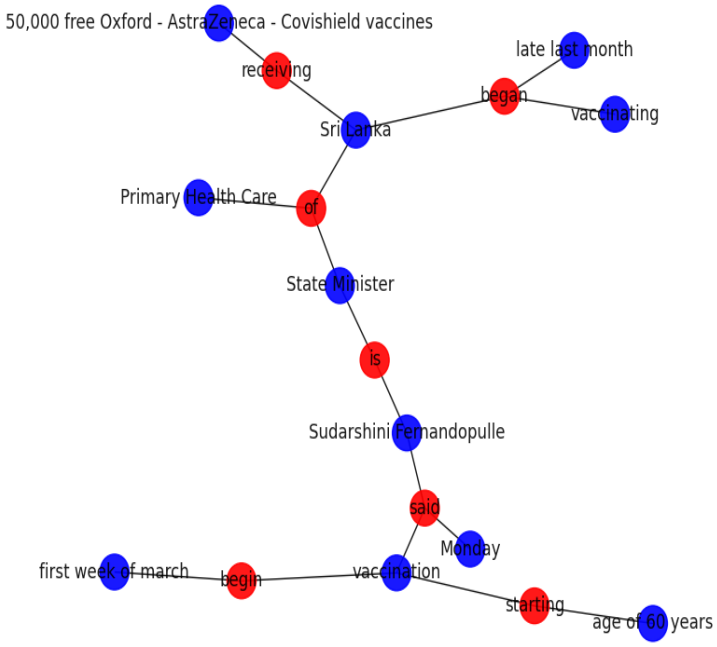


**Figure 8.** A portion of sample KG generated by our pipeline.

Entities are represented in blue color nodes, while relationships are in red color nodes. Another significant aspect of our KG pipeline is the hierarchal relations detection process. For example, the sentence "Countries such as Sri Lanka receiving free vaccines . . . " is shown in Figure 9.

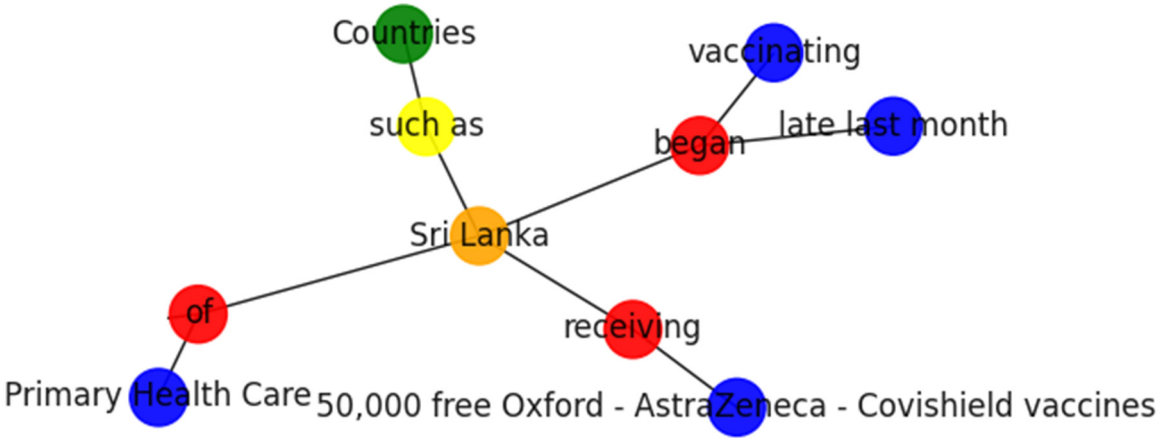


**Figure 9.** Representation of hierarchical relations using Hearst patterns.

Following the Hearst patterns in Table 1, the KG pipeline identifies the relation, such as Hypernym: Countries and Hyponym: Sri Lanka, as shown in Figure 9 using yellow, green, and orange color nodes, respectively.

This study uses a personal computer with the following configurations to build and evaluate the KG pipeline. The Intel suggestions would be an Intel(R) Core (TM) i3-8130U CPU @ 2.20 GHz or higher and at least 8 GB RAM along with a 1 TB HDD for performing better in KG generation.

## 5. Evaluation

We ran our KG pipeline on the user query "COVID-19 vaccination in Sri Lanka", and based on the e-news articles in the web search results, we crawled and prepared the e-news corpus using 10 news articles retrieved for the above user query. The distribution

of extracted e-news article lengths, measured by the number of sentences, is visualized in Figure 10.

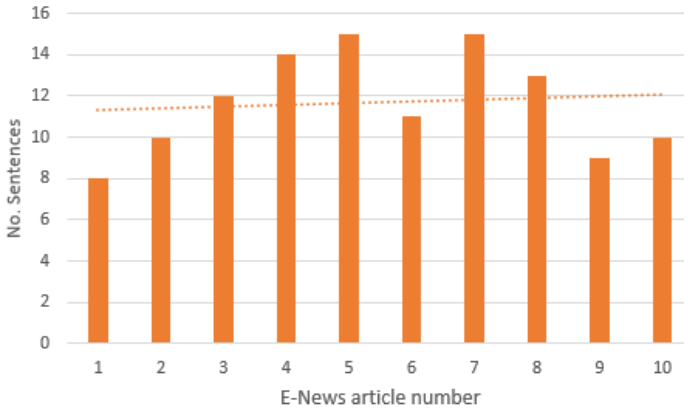


**Figure 10.** E-news article—length distribution.

According to the analysis, many of the e-news articles have sentences of fewer than 12, and e-news articles with more than 12 sentences are very infrequent. The distribution of Named Entity types in extracted e-news corpus is described in Figure 11.

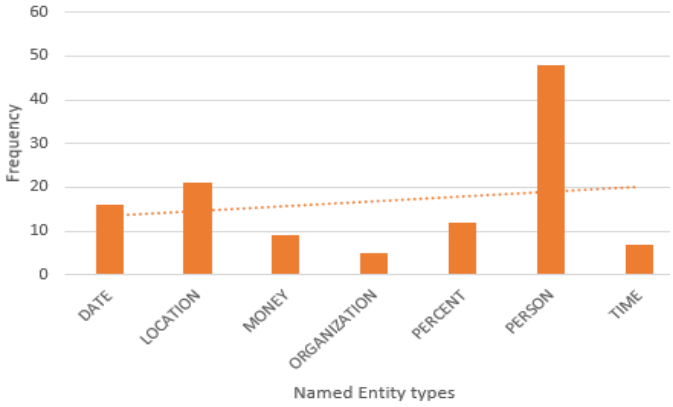


**Figure 11.** Named Entity type distribution.

As expected in our user query, the top three entity types are DATE, LOCATION, and PERSON, which is nearly around 72% of the unique named entities present in the extracted e-news corpus. Figure 12 shows the distribution of extracted triples from e-news articles in our corpus.

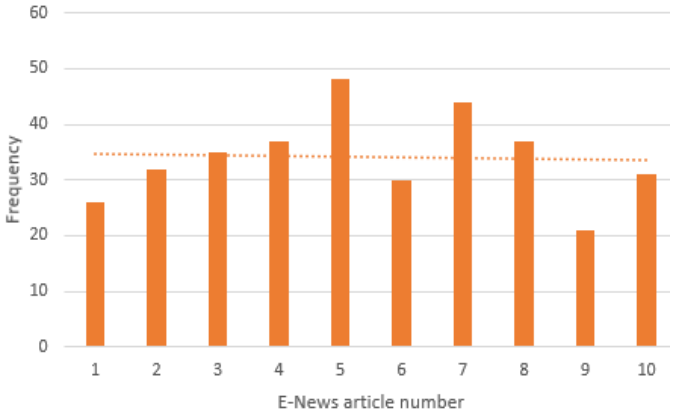


**Figure 12.** Triple distribution in KB.

A total of 341 triples constituting 96% of unique triples were extracted from 117 sentences included in 10 e-news articles. More than three-quarters of the triples were directly extracted via appositions. The Hearst-pattern-based relation extraction module and SRL module inferred 27 new noun/verb-mediated relations contributing to the growth of KB, while the SRL module detects over 7% new or missing links out of total relations.

Further, the authors conducted a study to examine the confidence of inferencing the information in existing and discovered relationships in the KB. As an intrinsic evaluation strategy, we use low-dimensional word embedding to show the effectiveness of the triples extracted by the KG pipeline. Two distinctive types of scoring functions, distance-based ($d_t$) scoring and similarity-based ($s_t$) scoring [10], are calculated using cosine similarity and used to measure the plausibility of triples in a KB. Mappings of words in the triples to vectors of real numbers was completed using the Word Embedding: Word2Vec model.

We calculate the distance-based ($d_t$) and similarity-based ($s_t$) scoring values for each triple in the KB, and the results of few triples are described in Table 2. As an example, the translational distance between h+r (Sri Lanka + began) and t (vaccination) is 0.0187 (1.87%), while the multiplicative formulation of semantic similarity of this triple is equal to 0.8999 (90%). According to the analysis, we observe 87% of triples in the KB have $d_t$ lesser than the average of 0.3861, which means that translational distance between h+r and t are lesser. Similarly, the semantic similarity of most of the triples in the KB is over the average of 0.8525 (85%), which means that the SPO triple has a significant semantic value among them.

**Table 2.** Results of scoring functions.

| Triple | $d_t$ | $s_t$ |
|---|---|---|
| (“Sri Lanka”, “began”, “vaccination”) | 0.0187 | 0.8999 |
| (“Sri Lanka”, “began”, “late last month”) | 0.0582 | 0.8902 |
| (“Sri Lanka”, “receiving”, “50,000 free Oxford-AstraZeneca-Covishield vaccines”) | 0.0232 | 0.9120 |

## 6. Discussion

Our KG pipeline was evaluated with 10 e-news articles retrieved from several e-news websites based on the user query “COVID-19 vaccination in Sri Lanka”. The challenge of extracting the only web news content while eliminating structured information and disambiguation pages was achieved using the Python-goose library. Furthermore, the pre-processing phase followed by the most required NLP techniques presents a significant trade-off in filtering the properties that further describe entities and relations. We emphasize only a limited type of entities that are important in the news domain, including DATE, LOCATION, MONEY, ORGANIZATION, PERCENT, PERSON, and TIME. Out of those entities, the majority of Named Entity types in extracted e-news corpus are DATE, LOCATION, and PERSON, which is nearly around 72% of the unique named entities present in the extracted e-news corpus. So, this study supports identifying the more frequent type of entities that are important in the news domain.

The rule-based SPO triple extraction approach in the Knowledge Base Construction phase is similar to the approach described by Huang et al. [2], and this was the most important factor for the precision in KB as it directly extracted a total of 341 triples constituting 96% of unique triples presented in the e-news corpus. This study highlights the significance of following the rule-based SPO triple extraction to extract nearly around 9% triples presented in the e-news corpus that cannot be detected without following the rule-based SPO triple extraction. Handling co-referential mentions is another noteworthy step in the Knowledge Base Construction phase. The outputs obtained from the co-reference resolution using the Stanford CoreNLP libraries were a significant factor for identifying all those expressions or phrases that denoted the same entity. Our KG pipeline combined all the identified co-referential mentions and thereby localized prominent contents related to each entity in the KG and as well as made inter-connections among spread-out entities

through graph paths. Both rule-based SPO triple extraction and co-reference resolution approaches cause our KG pipeline to outperform in KB generation.

Hearst-pattern-based methods described in Table 1 have been one of the most influential approaches for us to recognize the hierarchical level relationships in the e-news corpus. The Hearst-pattern-based relation extraction module in the KG pipeline inferred over five new noun/verb-mediated relations in the used e-news corpus. After establishing a steady KB, we followed the SRL frameworks elaborated in [3,12,14] to infer all the missing links in the KB and to identify newly inferred relational links, and add those to the existing KB. As per the evaluations, the resulting KG encompasses 22 new noun/verb-mediated relations while detecting over 7% new or missing links out of total relations based on the available observations that cannot be explicitly derived in addition to the original facts extracted at the beginning. Hence, this study highlights the significance of the SRL framework in inferring novel knowledge as triples without sacrificing precision.

Existing general-purpose KGs do not support building a scalable KG to represent the knowledge in a collection of e-news articles based on a given user query. Hence, personalized, scalable KGs [1,4,8] that use encyclopedic KGs, such as YAGO, DBPedia, and Freebase, are closely aligned with our objective of generating scalable KGs, but they do not support generating KGs using extracted web articles on the internet. Owing to this, it is hard to compare the resulting KG with existing KGs as their ontologies are different. Moreover, the evaluations in KGs still remain difficult due to the absence of a standardized set of evaluations and limitations in time and resources. Inspired by the study of Ji et al. [10], we evaluated the plausibility of each triple in the resulting KG using distance-based and similarity-based scoring functions. According to the analysis, we observe 87% of triples in the KB have a lesser translational distance between h+r and t, which means that head and tail entities have a strong association with the relation. Similarly, the semantic similarity of 83% of triples in the KB is over the average of 0.8525, which means that the SPO triple has a significant semantic value among them. Hence, our evaluation guarantees the capability of identifying how far triples in the KB would hold accurate information.

## 7. Conclusions and Future Works

This paper studies a novel KG construction pipeline to facilitate knowledge representation of a corpus of event-related e-news articles. All the explanations on how the entire KG pipeline works are provided in the research methodology.

The capability of identifying inaccurate triples or facts according to human perceptions is still an emerging research topic in the KG research community. Currently, there are multiple methods to compute the confidence of inferencing the information in existing and discovered relationships. Subsequently, the confidences in the inference are combined in the KG pipeline. Once the KG has been built, the conclusion on how far the facts will be considered to be true can be made. Nevertheless, this approach does not offer a confident method to verify that the fact will be assessed as a lawful fact by a human evaluator. So, the authors in this research carried out a detailed clarification on how to infer missing links using SRL frameworks and interrelate missing links into the KB to align the triples in the KG. We guarantee the detection of new or missing links using the SRL framework, with accounts of over 7% of total relations. Further, the distance-based and similarity-based scoring values are observed as worthy evidence of identifying the SPO in extracted triples that have occurred in similar contexts, which tend to be closer to each other in vector space.

Inspired by the discrepancy between the current large-scale KGs and the relatively limited information desires of e-newspaper readers, the authors in this research contributed to constructing a completely automated KG pipeline to generate a KG that can be used as an efficient application in analyzing and generating knowledge representation for a set of e-news articles for a particular news event based on a given user's queries. Inclusively, it has been observed that the scalable KG pipeline in this study generates a comprehensive and precise knowledge representation for the corpus of e-news articles.

The authors emphasized only a limited type of entities that are important in the news domain, including DATE, LOCATION, MONEY, ORGANIZATION, PERCENT, PERSON, and TIME during the Knowledge Base Construction due to the time limitations in incorporating all entity types in the KB. Furthermore, another limitation in this study is the sparsity in Hearst patterns to detect the words related to the Hearst patterns unless they co-occur in a sentence in exactly the correct configuration.

The authors of this study have identified many possibilities for future directions, including integration of word sense instantiation to enhance the accuracy of knowledge representation generated by KG. The integration of the rich semantics provided by ontologies such as WordNet ontology to establish different types of relations among existing nodes in the KG, such as synonymous relationships, is another direction to proceed in the future. This will increase the consistency and compactness of the KG and thereby perform a more sophisticated analysis and semantic normalization using the generated KG. As such, authors believe our current work unlocks potential avenues for emerging approaches in knowledge representations.

**Author Contributions:** M.V.P.T.L. designed the model, the computational framework, performed the experiments, and analyzed the data under the guidance of H.A.C., M.V.P.T.L. wrote and improved the manuscript in consultation with H.A.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The e-news articles presented in this study are available on https://www.hindustantimes.com/ (accessed on 20 March 2021).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Wu, P.; Zhou, Q.; Lei, Z.; Qiu, W.; Li, X. Template Oriented Text Summarization via Knowledge Graph. In Proceedings of the 2018 International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, China, 16–17 July 2018; pp. 79–83. [CrossRef]
2. Huang, L.; Wu, L.; Wang, L. Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 5094–5107. [CrossRef]
3. Elhammadi, S.; Lakshmanan, L.V.; Ng, R.; Simpson, M.; Huai, B.; Wang, Z.; Wang, L. A High Precision Pipeline for Financial Knowledge Graph Construction. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 12 December 2020; pp. 967–977. [CrossRef]
4. Faber, L.; Safavi, T.; Mottin, D.; Müller, E.; Koutra, D. Adaptive Personalized Knowledge Graph Summarization. In Proceedings of the 14th International KDD Workshop on Mining and Learning with Graphs (MLG), London, UK, 20 August 2018.
5. Hulliyah, K.; Kusuma, H.T. Application of knowledge graph for making Text Summarization (Analizing a text of educational issues. In Proceedings of the 3rd International Conference on Information and Communication Technology for the Moslem World (ICT4M) 2010, Jakarta, Indonesia, 13–14 December 2010; pp. E79–E83. [CrossRef]
6. Malviya, S.; Tiwary, U.S. Knowledge-Based Summarization and Document Generation using Bayesian Network. *Procedia Comput. Sci.* **2016**, *89*, 333–340. [CrossRef]
7. Song, Q.; Wu, Y.; Dong, X.L. Mining Summaries for Knowledge Graph Search. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 1887–1900. [CrossRef]
8. Safavi, T.; Belth, C.; Faber, L.; Mottin, D.; Muller, E.; Koutra, D. Personalized Knowledge Graph Summarization: From the Cloud to Your Pocket. In Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM), Beijing, China, 8–11 November 2019; pp. 528–537. [CrossRef]
9. Gunaratna, K. Semantics-based Entity Summarization. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, Melbourne, Australia, 19–25 August 2017; pp. 1060–1066.
10. Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; Yu, P.S. A Survey on Knowledge Graphs: Representation, Acquisition and Applications. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**. [CrossRef] [PubMed]
11. Wang, P.; Jiang, H.; Xu, J.; Zhang, Q. Knowledge Graph Construction and Applications for Web Search and Beyond. *Data Intell.* **2019**, *1*, 333–349. [CrossRef]
12. Al-Khatib, K.; Hou, Y.; Wachsmuth, H.; Jochim, C.; Bonin, F.; Stein, B. End-to-End Argumentation Knowledge Graph Construction. In Proceedings of the AAAI Conference Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 7367–7374. [CrossRef]

13. Nicholson, D.N.; Greene, C.S. Constructing knowledge graphs and their biomedical applications. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1414–1428. [CrossRef] [PubMed]
14. Conceptualizing the Knowledge Graph Construction Pipeline. Available online: https://towardsdatascience.com/conceptualizing-the-knowledge-graph-construction-pipeline-33edb25ab831 (accessed on 26 May 2021).
15. Roller, S.; Kiela, D.; Nickel, M. Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia, 15–20 July 2018; pp. 358–363. [CrossRef]
16. Atzori, M.; Balloccu, S. Fully-Unsupervised Embeddings-Based Hypernym Discovery. *Information* **2020**, *11*, 268. [CrossRef]
17. Wang, Q.; Mao, Z.; Wang, B.; Guo, L. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2724–2743. [CrossRef]