Computing and Informatics, Vol. 40, 2021, 318–340, doi: 10.31577/cai_2021_2_318

CLUSTERING AND BOOTSTRAPPING BASED FRAMEWORK FOR NEWS KNOWLEDGE BASE COMPLETION

K. Srinivasa, P. Santhi Thilagam

National Institute of Technology Karnataka
Department of Computer Science and Engineering
NH 66, Srinivasnagar, Surathkal, Mangalore
Karnataka – 575 025, India
e-mail: srinivas.karur@gmail.com, santhi@nitk.edu.in

Abstract. Extracting the facts, namely entities and relations, from unstructured sources is an essential step in any knowledge base construction. At the same time, it is also necessary to ensure the completeness of the knowledge base by incrementally extracting the new facts from various sources. To date, the knowledge base completion is studied as a problem of knowledge refinement where the missing facts are inferred by reasoning about the information already present in the knowledge base. However, facts missed while extracting the information from multilingual sources are ignored. Hence, this work proposed a generic framework for knowledge base completion to enrich a knowledge base of crime-related facts extracted from online news articles in the English language, with the facts extracted from low resourced Indian language Hindi news articles. Using the framework, information from any low-resourced language news articles can be extracted without using language-specific tools like POS tags and using an appropriate machine translation tool. To achieve this, a clustering algorithm is proposed, which explores the redundancy among the bilingual collection of news articles by representing the clusters with knowledge base facts unlike the existing Bag of Words representation. From each cluster, the facts extracted from English language articles are bootstrapped to extract the facts from comparable Hindi language articles. This way of bootstrapping within the cluster helps to identify the sentences from a low-resourced language that are enriched with new information related to the facts extracted from a high-resourced language like English. The empirical result shows that the proposed clustering algorithm produced more accurate and high-quality clusters for monolingual and cross-lingual facts, respectively. Experiments also proved that the proposed framework achieves a high recall rate in extracting the new facts from Hindi news articles.

Keywords: Knowledge base completion, natural language processing, information extraction, triples, bootstrap, cluster

1 INTRODUCTION

Knowledge Bases (KBs) contain a huge collection of information in the form of entities and relations extracted from structured and unstructured sources. Such information is stored as triples in machine-readable form like $\langle e_1 - R - e_2 \rangle$ called as facts. Knowledge Base Completion (KBC) is a long-standing problem in the area of knowledge management that involves the task of identifying the missing facts from the KBs. To date, the KBC problem is studied as a Knowledge Graph (KG) refinement problem where the missing facts are inferred from the existing facts in the KB [1]. For example, $works\_for$ relation can be inferred from the fact $\langle Person\_X - CEO\_of - Comapny\_Y \rangle$ by applying the appropriate inferring techniques. The focus of such techniques is on improving the inferring accuracy so that more number of appropriate hidden facts are extracted from the KB. Hence, these techniques ensure the identification of new facts that are not explicitly stored but are hidden in the KB. However, it is also necessary to ensure the completeness of the KB by identifying the missing facts while extracting the information from multiple sources.

In the era of a multilingual environment where the information is scattered across the web in multiple languages, most of the facts are redundant but are enriched with some new facts. For instance, the online news articles from different sources with various native languages within the same window of published dates, include information related to almost similar facts. However, each source may be enriched with some new facts about an event, and failing in identifying such facts is censorious for applications like crime prevention and monitoring. For instance, "Gurgaon police arrests key Lawrence Bishnoi gang member from Hyderabad" and लॉरेंस बिश्नोई गैंग का क़ुख्यात अपराधी संपत नेहरा हैदराबाद से गिरफ्तार, कई राज्यों में था इसका आतंक shows two sample headlines from English and Hindi news articles, respectively [35, 36]. Even though both the headlines contain information about the same event, the info who was arrested is missing from the Hindi language headline. Hence, for applications that develop KB from news articles, it is not sufficient to extract the information only from English news articles to ensure the completeness of KB.

In this paper, we extended the work proposed by [2] to develop a KB called "Crime Base" which was enriched with the facts extracted from only English news articles. KB so developed was proved to be incomplete by manually cross verifying the related bilingual English-Hindi language articles. However, the task of grouping the related articles across the languages and extracting the facts from articles in Indian languages like Hindi needs language-specific tools like Parts of Speech (PoS) tagger. However, these tools are either unavailable or not accurate enough to be

used for low-resourced languages like Indian languages. Although the grouping of articles can be achieved using document clustering techniques [3], the traditional way of representing the clusters using Bag of Words is not appropriate due to its inability to represent the cluster semantically. The semantic way of representing the cluster is highly essential as the quality of clusters formed depends on how semantically a cluster is represented. Besides, it is also essential to cluster the news articles incrementally as they are published daily and are to be treated as data streams. Hence, it is most appropriate to adopt the methods used to cluster the data streams to cluster the news articles [4]. Nevertheless, these methods are to be modified to cluster the articles across the languages. Even though open information extraction (OIE) tools like ArgOE are best suited to extract the information from multiple languages, they are developed to be used with foreign languages like English, Spanish, and Portuguese [5]. A straightforward way to solve the problem which is adopted by the existing works is to translate the entire document written in a target language like Hindi to source language English [6]. Such methods are time expensive as all the texts available in a target language do not contribute to the extraction of facts. In contrast, translation of only named entities like name of the PERSON, ORGANIZATION, and LOCATION and their relationships is adequate to compare to the translation of the whole document from KB perspective. In the news domain where the corpus is a collection of multilingual news articles, the extraction of facts from such a corpus is possible using supervised machine learning techniques. However, these techniques require a large collection of sententially aligned parallel data from different language articles to train the system which is very expensive.

The news articles are the kind of comparable corpora where the multilingual articles within a window of published dates are usually redundant. Extracting information from such a corpus can be achieved by clustering the articles across the languages based on their topical similarity. Once the topically similar articles are grouped, the source language facts can be bootstrapped to extract the facts from target language articles. Such a bootstrapped way of extraction limits the translation only to named entities and their relationships and hence reduces the time required to translate the entire article from the target language to the source language. Accordingly, this work proposes a bootstrapping-based KBC framework that can be adaptable to any domain and language using the appropriate machine translation tool. Moreover, the framework also helps to extract the facts from target language articles without using language-specific tools like POS tags which are most necessary for Indian languages. The overall architecture of the proposed work is shown in Figure 1.

The primary contributions of this paper are as follows:

1. Proposed an algorithm for grouping the related news articles in a bilingual corpus.
2. Proposed a bootstrapping-based method to extract the facts from target language news articles using facts extracted from source language news articles with minimum translation efforts.

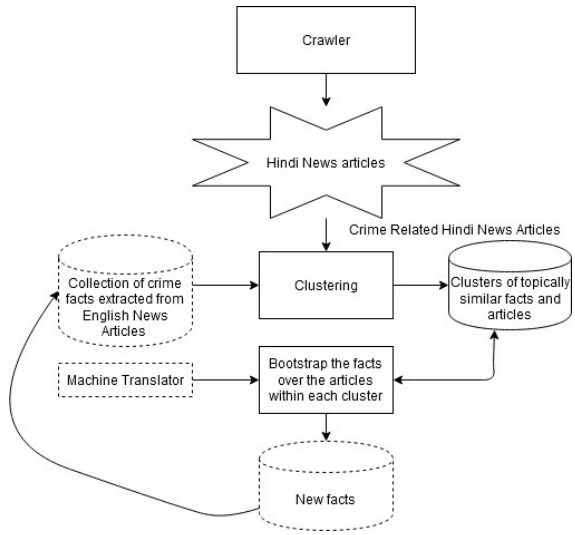

Figure 1. High level architecture of the proposed framework

The rest of the paper is organized as follows. Section 2 discusses the background and related work. Section 3 describes the problem along with the research objectives. The proposed methodology is explained in Section 4. Section 5 presents the results and analysis of the experiments conducted. Conclusion and future works are explained in Section 6.

2 BACKGROUND AND RELATED WORK

This section aims to provide an overview of the work in the domain of knowledge base construction with an emphasis on knowledge base completion. Knowledge base construction is the main vision of the semantic web to create a shared repository of KB in machine-readable form. Even though the problem of knowledge base construction is studied for decades, the knowledge base completion is studied only as a knowledge base refinement problem [1, 29, 30]. The knowledge base refinement methods try to complete the knowledge base by considering the facts internal to the knowledge base by inferring the new facts hidden inside the given knowledge base. However, these methods do not cover the external facts, i.e. facts extracted from multiple sources while constructing the knowledge base. In this perspective, the knowledge base completion can be treated as a problem of information extraction and integration, where the final knowledge base must be enriched with all the new facts extracted from multiple sources.

The existing works to extract the information from multiple sources are categorized as monolingual and multilingual based on the languages they support. Systems that extract the knowledge from sources in a single language are considered as mono-

lingual systems and most of the systems extract the knowledge only from sources in English language [7]. The authors in [8] extract the crime-related information from English news articles. [9] developed a system to construct a sports knowledge base by extracting the information from the FIFA website in the English language. [31] proposed an NLP and machine learning-based method for extraction of economic events and constructed a financial knowledge base. A T2KG system is proposed by [32] to construct a knowledge graph from unstructured text. An artist's knowledge base is constructed by [10] by extracting the information from the related websites in English. Few works also contributed to creating a knowledge base of events collected from news articles. For instance, Storybase was the knowledge base created by extracting the information from daily web news and Wikipedia current events [11]. In [12] authors extract the facts from French-language news articles. Knowledge Vault [13] is a probabilistic system that combines extractions from multiple sources like text and tabular data. PRISMATIC is a large-scale lexicalized relation resource that automatically extracts the knowledge from articles in English language [14]. Even though a lot of works are carried out to create a knowledge base by extracting the information from multiple sources, these works do not emphasize completing the knowledge base from multilingual sources.

In contrast to monolingual systems, systems that extract the knowledge from sources of different languages are considered as multilingual systems. Most of the multilingual systems consider the sources in foreign languages [15, 6, 16, 17, 18, 5] and only [19] extracts the Indian language Hindi along with foreign languages. However, these systems are based on either using language-specific processing tools or translating the entire documents into English.

Apart from the individual efforts in generating KBs like [20], few integrative projects which involve a community of users in creating KBs by extracting and updating the facts from crowd-sourced data like Wikipedia also emerged. To name few, Yago [21] is a KB created automatically from Wikipedia, WordNet and Geonames. DBpedia [22] exploits both free text as well as semi-structured data like infoboxes from Wikipedia to create the KB. BabelNet [23], the largest repository of multilingual words and senses, integrates Wikipedia and WordNet for creating the KB. Wikidata [24] is a KB enriched with facts extracted only from Wikipedia. Even though, the KB generated by these systems are well structured to support the web of linked data, the facts covered and validated by these systems are limited to Wikipedia.

There are some open-source tools like FRED and FOX [25] that generate structured knowledge graphs from unstructured texts. FRED is a powerful tool that extracts the knowledge from 48 languages. However, the capability of the tool is limited to only extraction and lack in integrating the knowledge extracted from multiple languages.

In addition to the efforts to generate the knowledge base, several studies attempted to develop knowledge base completion models using cross-lingual projection of knowledge. However, these models require the presence of a knowledge base for both the source and target language. Using the knowledge bases for both the

languages, the facts from the source language are projected with the target language for knowledge base completion. For instance, [26] and [27] developed a knowledge base completion model based on vector representation by representing the concepts in multiple languages in a unified vector space. But these models are not applicable in the absence of a knowledge base for a target language.

Table 1 shows the consolidated view of the features supported by existing works on knowledge base construction. The table lists both monolingual as well as multilingual systems. From the table, it is clear that none of the systems supports knowledge base completion by identifying the missing facts while extracting the information from multiple sources. Specific to the news domain, the existing systems considered only English news articles and ignored the facts available in other language news articles. Moreover, exploiting the redundancy that exists among the news articles for the identification of new facts is also underexplored which is observed from the last column in the table.

3 PROBLEM DESCRIPTION AND RESEARCH OBJECTIVES

Given a set of bilingual news articles from resource-rich Source Language (SL) like English and resource deficit Target Language (TL) like Hindi. This paper aimed at developing a framework for KBC to extract the facts from TL news articles so that the KB created using SL news articles is enriched with new facts available in TL news articles. This is to be achieved by exploiting the redundancies available from SL and TL news articles and without using the language-specific tools for TL news articles.

To address the problem described above, the following research objectives or tasks are set:

Task-1: To propose an algorithm for grouping the related articles from SL and TL using clustering.

Task-2: To propose a method to extract the facts from TL news articles using facts related to SL news articles as a bootstrapping data set and an appropriate machine translation tool.

4 METHODOLOGY

The detailed architecture of the proposed framework is shown in Figure 2. The proposed framework performs bootstrapping at multiple levels to extract the new facts from Hindi news articles using the triples extracted from the related English news articles. The framework consists of two main stages, namely clustering and extraction, to solve respective tasks mentioned in Section 3 and are discussed in the following sections.

System	Method of IE	Domain	Source of Extraction	Language	Integration	Support for KBC	Exploitation of Redundancy
CrimeProfiler ([8])	Stanford NER, Semi-supervised	Crime	News articles	English			
SOBA ([9])	SProUT, Rule based	Sports	FIFA website	English	✓		
Music Knowledge Base ([20])	Rule based, Un-supervised	Music	songfacts.com	English	✓		
Artequakt ([10])	Wornet, GATE, Ontology	Artists	Web	English	✓		
PRISMATIC ([14])	Frame+Un-supervised	Open	Web	English			
NELL ([15])	Semi-supervised	Open	Web	English	✓		
RdfLiveNews ([16])	Un-supervised, Supervised	Open	Web	English	✓		
Storybase ([11])	Un-supervised	Open News	Web	English			
Stern et al. ([12])	Entity based	Open News	Web	French			
Knowledge Vault ([13])	Distant Supervision	Open	Web	English	✓		
Finance KB ([31])	NLP, Machine Learning	Economic	News	English			
T2KG ([32])	Rule and similarity based	Open	Web	English			
Even centric knowledge grahs ([6])	Semantic Role Labelling	Open News	Wikinews, FIFA world cup, Global Automotive Industry, Airbus A380 airplanes.	English, Spanish, Italian and Dutch	✓		
BOA ([7])	Bootstrapping	Open	Web	English, German	✓		
New/s/leak ([17])	Using Polyglot tool	Open News	Web	40			
ZENON ([18])	GATE, Rule Based	Crime	Intelligence Reports	English, Deri	✓		
Crime base ([2])(Knowledge base enriched with facts from English news articles)	Rule based	Open News	Web	English	✓		
Proposed Work(Extension of Crime base)	Rule based	Open News	Web	English, Hindi	✓	✓	✓

Table 1. Existing vs. proposed system

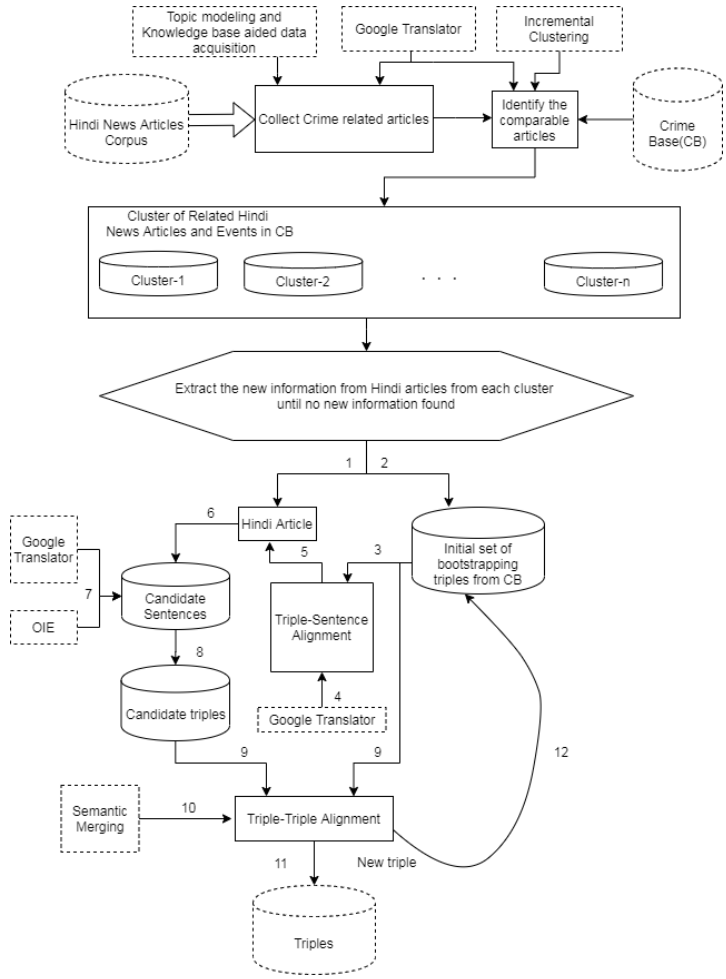

Figure 2. Detailed architecture of the proposed framework

4.1 Methodology for Task-1: Clustering

Initially, the crime-related Hindi articles are selected by applying topic modeling and knowledge base aided data acquisition method proposed by [2] over the headlines translated to English. The redundancies among the articles are exploited by identifying the comparable articles. A set of bilingual collections of articles is said to be comparable if they are related either topically or sententially. Topically related articles are contextually similar articles that discuss the same topic and are said to be semantically related. Whereas, sententially related articles are almost bilingual translations of each other and are said to be semantically similar. Hence topically re-

lated articles are enriched with more new information compare to sententially related articles. The proposed work identifies the comparable articles using the semantic merging procedure mentioned in [2]. As the news articles are published daily, the articles are considered as data streams and an incremental nearest neighborhood algorithm for clustering data streams is adopted [4]. Here, the clustering algorithm is modified to identify the sententially and topically related articles by finding sentential and topical neighbors and is named as Sentential-Topical-Nearest-Neighborhood (STNN) algorithm which is described in Algorithm 1.

The major difficulty in clustering articles is in semantically representing the articles so that clusters of better quality can be formed. Due to the availability of a large number of terms as document features, the Bag of Words way of representing the documents does not capture the semantics hidden in the sentences. To improve the semantics, the articles are represented as KB facts in the form of triples extracted over the headlines. Accordingly, the headlines from Hindi news articles are translated to English, and facts from the translated headlines are extracted using the method proposed in [2]. When a stream of facts from English and Hindi news articles comes in, we divide them into various windows based on their date of publication. Now, events in the first window are clustered using neighborhood-based clustering. The similarity between each of the elements in the first window is calculated using contextual as well as semantic similarity measures. The significance of using both the similarity measures is empirically proved and can be found in [2]. Two elements are considered to be topically neighbors if their contextual similarity is greater than a threshold value. Such neighbors are also checked for their semantic similarity. If the semantic similarity is greater than their contextual similarity score, they form a separate cluster and will be added to the set of sententially similar clusters. Otherwise, they will be added to the set of topically similar clusters. If the contextual similarity score for any two elements is less than the threshold, the elements are independent and form two separate clusters. To represent a cluster, we find the medoid of each cluster, where the medoid is an element that has the maximum similarity with all other elements in the cluster. This limits further comparison between the medoids rather than with all the elements in the cluster. A similar method is followed to find the clusters for other windows. For each new cluster, we find from the former clusters the most similar cluster to them by calculating the similarity of the medoid event of the former clusters and the medoid of the new cluster. Based on their similarity, two clusters are merged and the medoid will be updated.

4.2 Methodology for Task-2: Extraction

In this work, we propose a method to identify and extract the new facts from a target language news article like Hindi using the facts extracted from related English news articles. This is achieved by bootstrapping the triples extracted from English news articles to identify the presence of related triples from comparable Hindi news articles. The proposed extraction method constitutes two steps, namely:

Algorithm 1: STNN Algorithm

Input: $E = \{F_{E_1}, F_{E_2}, \ldots, F_{E_m}\}$:Set of m crime facts extracted from English news articles and $H_H = \{H_1, H_2, \ldots, H_n\}$:Set of n headlines extracted from Hindi language news articles

Result: Set of n_s sententially similar clusters $C_s = \{C_1, C_2, \ldots, C_{n_s}\}$, Set of n_t topically similar clusters $C_t = \{C_1, C_2, \ldots, C_{n_t}\}$

1 Translate the headlines in Hindi to English and the set of translated headlines be $H_{H_T} = \{H_{t_1}, H_{t_2}, \ldots, H_{t_n}\}$
2 Extract the facts from H_{H_T} and let $H = \{F_{H_1}, F_{H_2}, \ldots, F_{H_k}\}$ be a set of k facts related to Hindi headlines
3 Divide the events from E and H into multiple windows $W = \{w_1, w_2, \ldots\}$ where $w_i \subseteq E \cup H$ indicates facts extracted from the articles published during i^{th} date
4 Find the neighbors and hence clusters for w_1 as follows:
5 Calculate contextual similarity CS between each new couple of elements F_{E_i} and F_{H_j}.
6 If $CS >$a threshold t_c then
7 Calculate semantic similarity SS between each couple.
8 If $SS >$ a threshold t_s then
9 the elements are sententially neighbors. Each set of neighbors represent a cluster and will be added to C_s.
10 Otherwise
11 the elements are topically neighbors. Each set of neighbors represent a cluster and will be added to C_t.
12 Otherwise
13 Add the elements to C_t as new clusters.
14 Find medoid of each cluster where, medoid is the element which has the maximum similarity with all the elements in the cluster.
15 Similarly find the neighbors and hence the clusters for the subsequent windows.
16 Calculate new clusters medoids.
17 Calculate the similarity between new medoids and medoids of old clusters.
18 If found a pair of contextually or semantically similar medoids
19 Merge the clusers.
20 Update medoid.
21 Add the merged cluster to the appropriate set.
22 Otherwise
23 Retain the clusters as it is.

1. candidate sentence identification,
2. new triple generation.

Each of the steps is explained in the following subsections.

4.2.1 Candidate Sentence Identification

From each cluster, the events related to English news articles are selected as an initial set of bootstrapping triples. Each of these triples is translated to the target language using Google translator API and used to query the Hindi articles to identify a set of sentences that are enriched with new facts and are called candidate sentences. Given a set of bootstrapped triples from English articles $B_E = \{t_{E_1}, t_{E_2}, \ldots, t_{E_n}\}$, a set of candidate sentences from Hindi articles $S = \{s_1, s_2, \ldots, s_m\}$ are obtained by aligning the sentences with the triples. Formally, a sentence s_i is said to be aligned with t_{E_j}, if an element e_k belongs to t_{E_j} is a substring of s_i. Finally, a sentence that constitutes the un-aligned part in it is selected as the candidate sentence. Otherwise it is considered as similar to t_{E_j}. Figure 3 illustrates the generation of candidate sentences with an example.

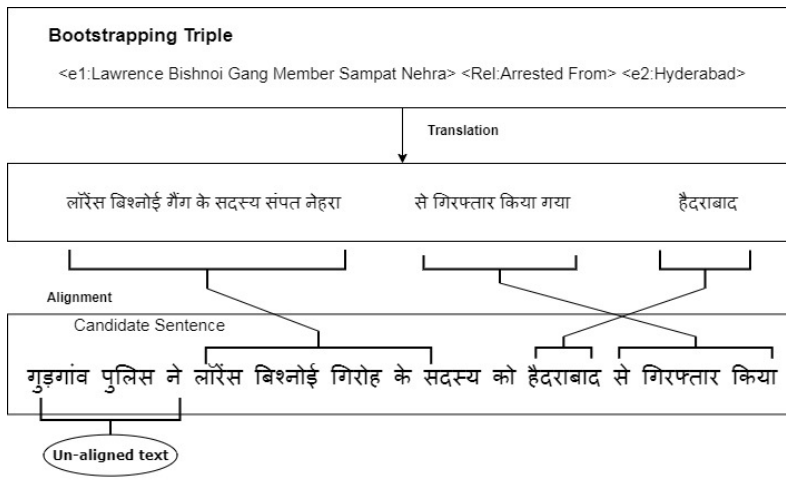

Figure 3. Candidate sentence generation

4.2.2 New Triple Generation

Once the candidate sentences are extracted, new triples are obtained in three steps, namely:

1. candidate sentence translation,
2. triple/s extraction,

3. projection of triple.

Initially, candidate sentences are translated to English language using Google API translator, and triples from each sentence are extracted using the method proposed in [2]. Triples so extracted from a candidate sentence are projected against the bootstrapped triple to identify the new triples, as shown in Figure 4 with the continuation of the example considered in Figure 3.

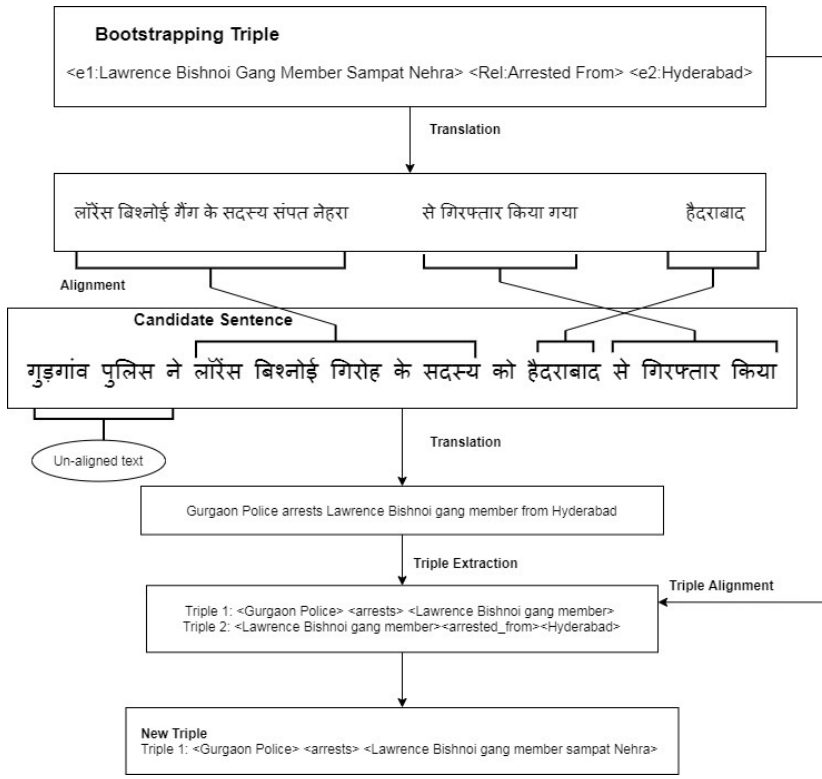

Figure 4. New triple generation

5 EXPERIMENTAL RESULTS

This work considers two prominent newspapers, namely *Indian Express* for English News articles and *Hindustan (हिंदुस्तान)*, which have articles available online. The corpus includes the data collected from *Jan* 2018 to *Jun* 2018. The following sections describe the experimental evaluation results for clustering and extraction.

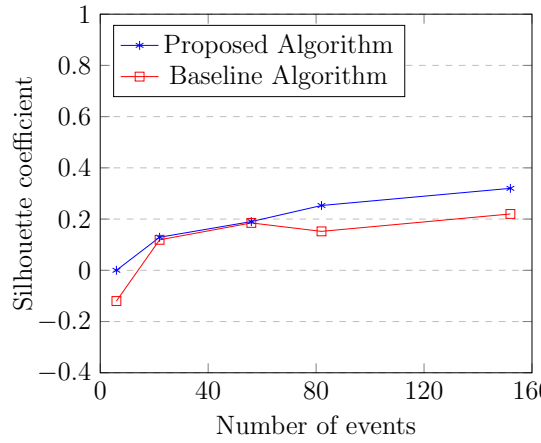

Figure 5. Bilingual evaluation of proposed clustering algorithm: Silhouette coefficient for varying number of events

5.1 Evaluation of Task-1: Clustering Algorithm

The proposed algorithm for clustering is evaluated in two phases, namely, bilingual and monolingual evaluation. In the first phase, the algorithm is evaluated for English and Hindi articles and in the second phase, the algorithm is evaluated for English articles. Due to the lack of algorithms for clustering multilingual articles, a baseline algorithm, i.e. incremental nearest neighborhood algorithm without using background KB and considering only the headlines from English and Hindi news articles, is implemented.

5.1.1 Phase-1 Evaluation: Bilingual Evaluation

The proposed algorithm is compared with the baseline algorithm in terms of the quality of clusters formed and the time taken for clustering. The clustering quality is determined using a Silhouette coefficient [28]. This is a well-known measure of internal evaluation for evaluating clusters without pre-determined labels. It measures how similar an object is to its cluster compared to other clusters. The Silhouette coefficient for i^{th} event is calculated as follows:

$$s_i = \frac{a_i - b_i}{\max(a_i, b_i)}$$

where a_i is the average similarity of the i^{th} event with all the other events in its cluster. Then for all the other clusters to which i^{th} event does not belong, we calculate the average similarity of i^{th} event to all the events in these clusters and b_i is the maximum of all these values. Figure 5 shows the silhouette coefficient obtained for proposed and baseline algorithms for varying numbers of events. We can see

that the proposed algorithm achieved a larger value of silhouette coefficient as the event size increases and hence produced a better quality of clusters. However, the Silhouette coefficient value is not very close to 1 because of many individual clusters obtained during the clustering process. These events are those which do not have similarity with any other crime events.

For instance, Figure 6 shows clustering results for 152 events extracted from 60 headlines using the proposed algorithm. There are many individual clusters and also clusters with a varying number of event elements. The medoid of each cluster can be seen highlighted in Figure 6. If we do not consider the individual clusters, then we get an average value of 0.63 and 0.45 as silhouette coefficients for proposed and baseline algorithms, respectively.

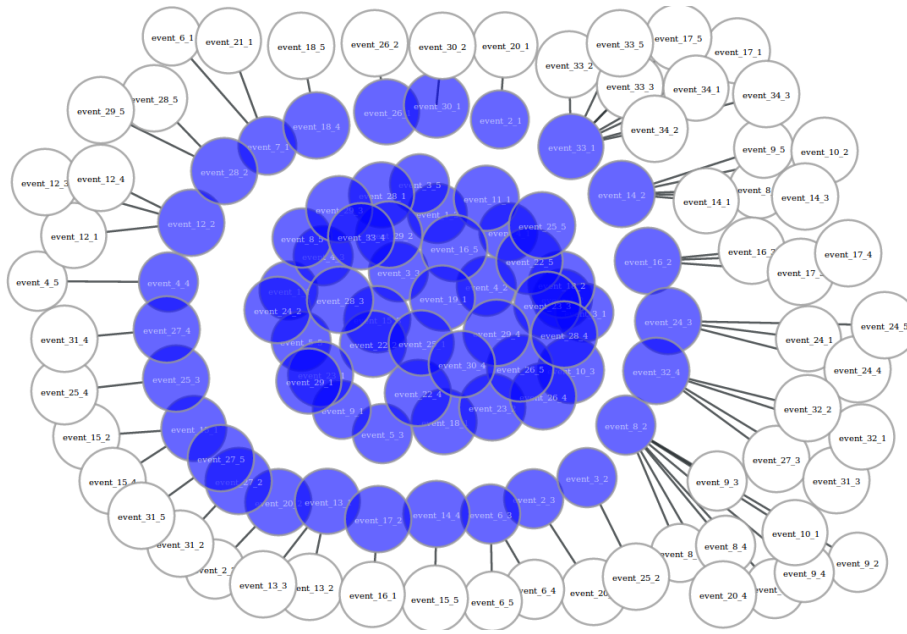

Figure 6. Visualization of cluster for 152 events

We also evaluated the quality of clusters in terms of the number of related events obtained for a given keyword. For instance, Figure 7 shows the clusters retrieved for the keyword "Navsari". It has 2 clusters associated with it with each cluster having 1 event. The keyword is directly related to both these events. The other attributes of the events are also shown. Some of the input keywords used for finding clusters of related events over a cluster in Figure 6 are shown in Table 2. From the table, it is clear that, due to the higher quality of clusters formed by the proposed algorithm, the number of related events associated with a given keyword is also significantly high.

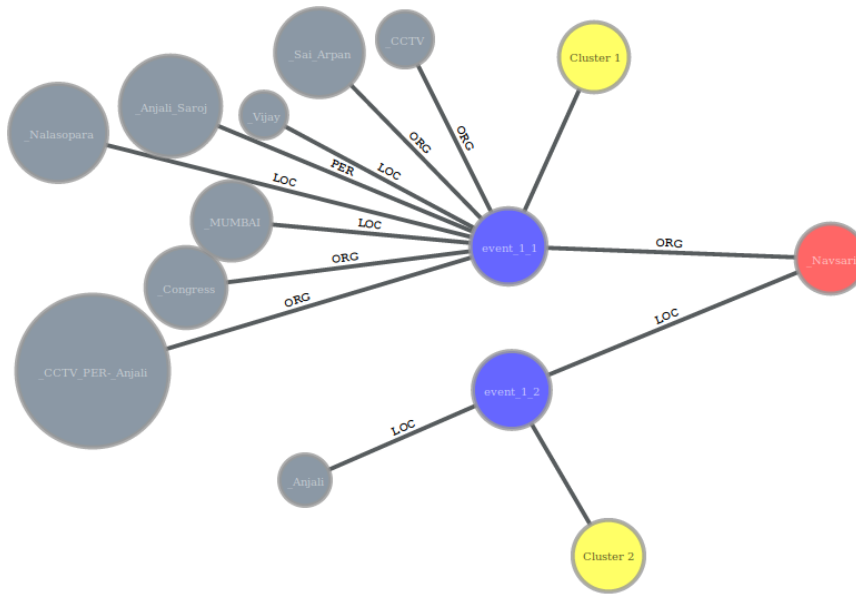

Figure 7. Cluster results for the keyword "Navsari"

Table 3 shows the clustering time taken by the proposed and baseline algorithms. From the table, it can be observed that the proposed algorithm takes more time compared to the baseline approach. This is due to the extraction of two or more triples from a single headline which is evident from the third column of the table. The time taken for machine translation and triple extraction are not considered for evaluation. However, more semantics hidden in triples compared to raw sentences

Keyword	Number of Related Events (Baseline Algorithm)	Number of Related Events (Proposed Algorithm)
Kanpur	1	3
Navsari	2	2
Venugopal	3	4
Malad	1	3
Mumbai	6	14
Bandipora	2	2
CRPF	7	9
Railway_Act	1	1
Abhijit Mukherjee	1	3
Kaluram	6	6
Congress	15	26

Table 2. Number of related events for keywords before and after clustering

produces clusters with high quality, and hence the time complexity is compromised over the cluster quality.

Features (Bag of Words in Terms of Headlines)	Clustering Time for Baseline Algorithm [s]	Features (Events in Terms of Triples)	Clustering Time for Proposed Algorithm [s]
100	92	270	98
200	194	423	222
300	298	610	343
400	372	908	402
500	536	1 022	582

Table 3. Bilingual evaluation of proposed clustering algorithm: Time taken for clustering

5.1.2 Phase-2 Evaluation: Monolingual Evaluation

Here the proposed work is evaluated by considering only the English news articles and comparing the results with two recently proposed works [33] and [34] as a baseline. Evaluation in these two works is done using Reuters and 20Newsgroup datasets. The details about the datasets can be found in [34]. [34] uses a K-means clustering algorithm with improved square root similarity measure. As an improvement to this, [33] used N-grams representation along with K-means clustering algorithm and improved square root similarity measure. The proposed algorithm is different from the baseline works by using semantically rich triples representation and a similarity measure using both contextual and semantic similarity measures proposed in [2]. Here, the experiment is conducted using 2000 samples each from Reuters and 20Newsgroup datasets over 5 newsgroups. The triples are extracted from each sample using the method proposed in [2]. To speed up the execution, a parallel version of the proposed clustering algorithm is implemented using Message Passing Interface (MPI). The triples are processed in parallel to identify the clusters.

Table 4 shows the evaluation results for the proposed and the baseline approaches. The same performance metrics as mentioned and defined in [33], i.e. accuracy and purity, are used here for evaluation. From the table, it is clear that the proposed clustering algorithm performs better than baseline methods in terms of accuracy. However, due to the generation of more individual clusters, i.e. clusters with a single element, the purity of the proposed algorithm is less compared to [33].

5.2 Evaluation of Task-2: Extraction

To evaluate the results for the proposed KBC approach, an MT-based system is implemented which is considered as a gold standard. The gold standard system reduces the problem to monolingual information extraction and integration by translating the entire articles in the target language into English. Then the facts are extracted

Methods	Datasets	Accuracy	Purity
[33]	Reuters	0.3950	0.9418
	20 Newsgroups	0.3801	0.9200
[34]	Reuters	0.2320	0.5769
	20 Newsgroups	0.1659	0.4234
Proposed Algorithm	Reuters	0.5210	0.6200
	20 Newsgroups	0.4832	0.7398

Table 4. Monolingual evaluation of proposed clustering algorithm

from translated articles and are semantically merged with facts extracted from English news articles using the methods for IE and semantic merging proposed by [2]. Hence the gold standard system is named as Machine Translation based Monolingual Knowledge Base Completion (MTML_KBC). The quality of the proposed KBC approach is measured using the standard evaluation metrics precision and recall. Precision is calculated as the ratio of the number of valid new facts extracted to the total number of new facts extracted. The recall is calculated as the ratio of the number of valid new facts extracted to the total number of valid new facts available.

Table 5 shows the results recorded for five different clusters. Figures 8 and 9 show the performance of gold standard (MTML_KBC) and proposed approach in terms of precision and recall, respectively. From the figures, it is clear that the proposed approach achieves a better recall compared to precision. This is evident from the fact that the total number of new facts extracted by the proposed approach is more due to improper projection of bootstrapping triples with the triples extracted from candidate sentences.

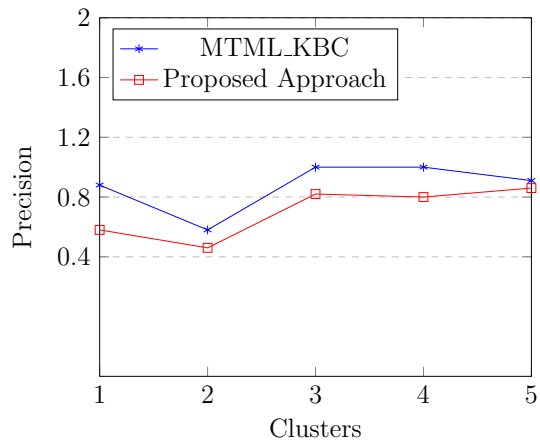

Figure 8. Precision for five clusters

Clusters	MTML_KBC			Proposed Approach		
	Total New Facts Extracted	Number of New Facts Available	Number of Valid New Facts Extracted	Total New Facts Extracted	Number of New Facts Available	Number of Valid New Facts Extracted
Cluster-1 (52 facts + 13 Hindi articles)	9	10	8	12	10	7
Cluster-2 (83 facts + 08 Hindi articles)	12	9	7	15	9	7
Cluster-3 (75 facts + 18 Hindi articles)	14	15	14	17	15	14
Cluster-4 (92 facts + 11 Hindi articles)	18	20	18	21	20	17
Cluster-5 (88 facts + 14 Hindi articles)	23	25	21	23	25	20

Table 5. Comparison of MTML_KBC and proposed approach

6 CONCLUSIONS AND FUTURE WORK

This work proposed a clustering and bootstrapping-based generic framework for knowledge base completion. Using the framework, any knowledge base created with the facts extracted from English news articles can be enriched with new facts available in low-resourced language articles without using language-specific tools. Here the experiment is conducted using the low resourced Indian language Hindi news articles. The redundancies that exist among the bilingual collection of articles are exploited by grouping the articles that are topically or sententially similar using the nearest neighborhood clustering. The proposed clustering algorithm makes use of knowledge base facts in terms of triples to represent the articles against the traditional Bag of Words representation, as the triples capture the high semantics. Empirical results show that clusters of high accuracy and quality are obtained for monolingual and bilingual facts, respectively. From each group of related articles, the facts related to English news articles are bootstrapped to extract the facts from Hindi news articles using Google translator API. This way of using the high-resource language facts as bootstrapping triples helps to extract the facts from articles related to the languages for which language processing tools like POS tags are neither available nor accurate. Experimental results for extraction show that using the framework a better recall is achieved in identifying the new facts compared to precision. A precision of high rate can be achieved by aligning the bootstrapped triples

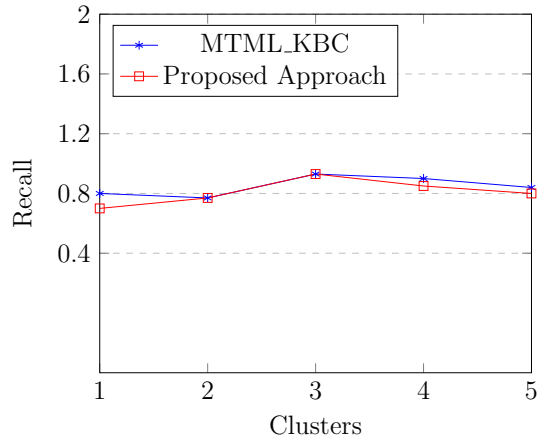

Figure 9. Recall for five clusters

with triples extracted from other languages more accurately, which will be considered in the future. In the future, the framework will be examined for other Indian languages also.

REFERENCES

[1] Paulheim, H.: Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. Semantic Web, Vol. 8, 2017, No. 3, pp. 489–508, doi: 10.3233/sw-160218.

[2] Srinivasa, K.—Thilagam, P. S.: Crime Base: Towards Building a Knowledge Base for Crime Entities and Their Relationships from Online News Papers. Information Processing and Management, Vol. 56, 2019, No. 6, Art. No. 102059, doi: 10.1016/j.ipm.2019.102059.

[3] Elsayed, A.—Mokhtar, H. M. O.—Ismail, O. : Ontology Based Document Clustering Using MapReduce. International Journal of Database Management Systems, Vol. 7, 2015, No. 2, doi: 10.5121/ijdms.2015.7201.

[4] Louhi, I.—Boudjeloud-Assala, L.—Tamisier, T.: Incremental Nearest Neighborhood Graph for Data Stream Clustering. 2016 International Joint Conference on Neural Networks (IJCNN'16), Vancouver, Canada, 2016, pp. 2468–2475, doi: 10.1109/ijcnn.2016.7727506.

[5] Claro, D. B.—Souza, M.—Castellã Xavier, C.—Oliveira, L.: Multilingual Open Information Extraction: Challenges and Opportunities. Information, Vol. 10, 2019, No. 7, Art. No. 228, 25 pp., doi: 10.3390/info10070228.

[6] Rospocher, M.—van Erp, M.—Vossen, P.—Fokkens, A.—Aldabe, I.—Rigau, G.—Soroa, A.—Ploeger, T.—Bogaard, T.: Building Event-Centric

Knowledge Graphs from News. Journal of Web Semantics, Vol. 37–38, 2016, pp. 132–151, doi: 10.1016/j.websem.2015.12.004.

[7] Gerber, D.—Ngomo, A.-C. N.: Extracting Multilingual Natural-Language Patterns for RDF Predicates. In: ten Teije, A. et al. (Eds.): Knowledge Engineering and Knowledge Management (EKAW 2012). Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 7603, 2012, pp. 87–96, ISBN: 978-3-642-33876-2, doi: 10.1007/978-3-642-33876-2_10.

[8] Dasgupta, T.—Naskar, A.—Saha, R.—Dey, L.: CrimeProfiler: Crime Information Extraction and Visualization from News Media. Proceedings of the International Conference on Web Intelligence (WI '17), 2017, pp. 541–549, doi: 10.1145/3106426.3106476.

[9] Buitelaar, P.—Cimiano, P.—Frank, A.—Hartung, M.—Racioppa, S.: Ontology-Based Information Extraction and Integration from Heterogeneous Data Sources. International Journal of Human-Computer Studies, Vol. 66, 2008, No. 11, pp. 759–788, doi: 10.1016/j.ijhcs.2008.07.007.

[10] Alani, H.—Kim, S.—Millard, D. E.—Weal, M. J.—Lewis, P. H.—Hall, W.—Shadbolt, N. R.: Automatic Extraction of Knowledge from Web Documents. 2nd International Semantic Web Conference – Workshop on Human Language Technology for the Semantic Web and Web Services, Sanibel Island, Florida, USA, 2003. Available at: https://eprints.soton.ac.uk/258194/.

[11] Wu, Z.—Liang, C.—Giles, C. L.: Storybase: Towards Building a Knowledge Base for News Events. In: Chen, H. H., Markert, K. (Eds.): Proceedings of ACL-IJCNLP 2015 System Demonstrations. ACL, 2015, pp. 133–138, doi: 10.3115/v1/p15-4023.

[12] Stern, R.—Sagot, B.: Population of a Knowledge Base for News Metadata from Unstructured Text and Web Data. Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction (AKBC-WEKEX), Montréal, Canada, ACL, 2012, pp. 35–40. Available at: https://aclanthology.org/W12-30.pdf.

[13] Dong, X.—Gabrilovich, E.—Heitz, G.—Horn, W.—Lao, N.—Murphy, K.—Strohmann, T.—Sun, S.—Zhang, W.: Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 601–610, doi: 10.1145/2623330.2623623.

[14] Fan, J.—Kalyanpur, A.—Gondek, D. C.—Ferrucci, D. A.: Automatic Knowledge Extraction from Documents. IBM Journal of Research and Development, Vol. 56, 2012, No. 3-4, pp. 5:1–5:10, doi: 10.1147/jrd.2012.2186519.

[15] Carlson, A.—Betteridge, J.—Kisiel, B.—Settles, B.—Hruschka, E. R.—Mitchell, T. M.: Toward an Architecture for Never-Ending Language Learning. Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI '10), Atlanta, Georgia, 2010, pp. 1306–1313.

[16] Gerber, D.—Hellmann, S.—Bühmann, L.—Soru, T.—Usbeck, R.—Ngomo, A.-C. N.: Real-Time RDF Extraction from Unstructured Data Streams. In: Alani, H. et al. (Eds.): The Semantic Web – ISWC 2013. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 8218, 2013, pp. 135–150, doi: 10.1007/978-3-642-41335-3_9.

[17] Wiedemann, G.—Yimam, S. M.—Biemann, C.: A Multilingual Information Extraction Pipeline for Investigative Journalism. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, Belgium, 2018, pp. 78–83, doi: 10.18653/v1/D18-2014.

[18] Hecking, M.—Schwerdt, C.: Multilingual Information Extraction for Intelligence Purposes. 13th International Command and Control Research and Technology Symposium (ICCRTS): "C2 for Complex Endeavors", Seattle, WA, 2008. Available at: http://dodccrp.org/events/13th_iccrts_2008/CD/html/papers/025.pdf.

[19] Akbik, A.—Chiticariu, L.—Danilevsky, M.—Kbrom, Y.—Li, Y.—Zhu, H.: Multilingual Information Extraction with PolyglotIE. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, Osaka, Japan, 2016, pp. 268–272. Available at: https://aclanthology.org/C16-2056.pdf.

[20] Oramas, S.—Espinosa-Anke, L.—Sordo, M.—Saggion, H.—Serra, X.: Information Extraction for Knowledge Base Construction in the Music Domain. Data and Knowledge Engineering, Vol. 106, 2016, pp. 70–83, doi: 10.1016/j.datak.2016.06.001.

[21] Rebele, T.—Suchanek, F.—Hoffart, J.—Biega, J.—Kuzey, E.—Weikum, G.: YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. In: Groth, P. et al. (Eds.): The Semantic Web – ISWC 2016. Springer, Cham, Lecture Notes in Computer Science, Vol. 9982, 2016, pp. 177–185, doi: 10.1007/978-3-319-46547-0_19.

[22] Lehmann, J.—Isele, R.—Jakob, M.—Jentzsch, A.—Kontokostas, D.—Mendes, P. N.—Hellmann, S.—Morsey, M.—van Kleef, P.—Auer, S.—Bizer, C.: DBpedia – A Large-Scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web, Vol. 6, 2015, No. 2, pp. 167–195, doi: 10.3233/sw-140134.

[23] Navigli, R.—Ponzetto, S. P.: BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, Vol. 193, 2012, pp. 217–250, doi: 10.1016/j.artint.2012.07.001.

[24] Erxleben, F.—Günther, M.—Krötzsch, M.—Mendez, J.—Vrandečić, D.: Introducing Wikidata to the Linked Data Web. In: Mika, P. et al. (Eds.): The Semantic Web – ISWC 2014. Springer, Cham, Lecture Notes in Computer Science, Vol. 8796, 2014, pp. 50–65, doi: 10.1007/978-3-319-11964-9_4.

[25] Gangemi, A.—Presutti, V.—Reforgiato Recupero, D.—Nuzzolese, A. G.—Draicchio, F.—Mongiovì, M.: Semantic Web Machine Reading with FRED. Semantic Web, Vol. 8, 2017, No. 6, pp. 873–893, doi: 10.3233/sw-160240.

[26] Chen, M.—Tian, Y.—Yang, M.—Zaniolo, C.: Multilingual Knowledge Graph Embeddings for Cross-Lingual Knowledge Alignment. Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17), Melbourne, Australia, 2017, pp. 1511–1517, doi: 10.24963/ijcai.2017/209.

[27] Klein, P.—Ponzetto, S. P.—Glavaš, G.: Improving Neural Knowledge Base Completion with Cross-Lingual Projections. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, 2017, pp. 516–522, doi: 10.18653/v1/e17-2083.

[28] Rousseeuw, P. J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, Vol. 20, 1987, pp. 53–65, doi: 10.1016/0377-0427(87)90125-7.

[29] Malaviya, C.—Bhagavatula, C.—Bosselut, A.—Choi, Y.: Commonsense Knowledge Base Completion with Structural and Semantic Context. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, No. 03: AAAI-20 Technical Tracks 3, pp. 2925–2933, doi: 10.1609/aaai.v34i03.5684.

[30] Pezeshkpour, P.—Tian, Y.—Singh, S.: Revisiting Evaluation of Knowledge Base Completion Models. Automated Knowledge Base Construction (AKBC 2020), 2020, doi: 10.24432/C53S3W.

[31] Benetka, J. R.—Balog, K.—Norvag, K.: Towards Building a Knowledge Base of Monetary Transactions from a News Collection. 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2017, pp. 1–10, doi: 10.1109/jcdl.2017.7991575.

[32] Kertkeidkachorn, N.—Ichise, R.: T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text. AAAI Workshops 2017, San Francisco, CA, USA, 2017.

[33] Bisandu, D. B.—Prasad, R.—Liman, M. M.: Clustering News Articles Using Efficient Similarity Measure and N-Grams. International Journal of Knowledge Engineering and Data Mining, Vol. 5, 2018, No. 4, pp. 333–348, doi: 10.1504/IJKEDM.2018.095525.

[34] Sohangir, S.—Wang, D.: Improved SQRT-Cosine Similarity Measurement. Journal of Big Data, Vol. 4, 2017, No. 1, Art. No. 25, 13 pp., doi: 10.1186/s40537-017-0083-6.

[35] English News Article. Available at: `https://indianexpress.com/article/india/gurgaon-police-arrests-key-lawrence-bishnoi-gang-member-sampat-nehra-underworld-gangster-5207714/`.

[36] Hindi News Article. Available at: `https://www.livehindustan.com/ncr/story-lawrence-bishnoi-gang-gangster-sampath-nehra-arrested-from-hyderabad-2000546.html`.

K. Srinivasa received his B.Eng. degree in computer science and engineering in 2004 from the Vijayanagara Engineering College, Bellary, Visvesvaraya Technological University, Belgaum, India and his M.Tech. degree in computer science and engineering from the National Institute of Technology Karnataka (NITK), Surathkal, India, in 2010, and he has been pursuing his Ph.D. degree in the Department of Computer Science and Engineering, at the National Institute of Technology Karnataka (NITK), Surathkal, India, from 2017. Since 2005, he has been with the Department of Computer Science and Engineering at Siddaganga Institute of Technology, Tumakuru, Karnataka, India where he is currently Assistant Professor. His current research interests include information extraction, natural language processing and knowledge management.

P. Santhi Thilagam received her B.Eng. degree in computer science and engineering in 1991, and the M.Eng. degree in computer science and engineering from College of Engineering, Guindy, Anna University, Chennai, India, in 1999, and her Ph.D. degree in information technology from the National Institute of Technology Karnataka (NITK), Surathkal, India, in 2008. Since 1996, she has been with the Department of Computer Science and Engineering at NITK Surathkal, where she is Professor. Her current research interests include database security, data management, data analysis, and distributed computing. She is Member of several technical associations, scientific committees and editorial boards. She was the recipient of the best Ph.D. thesis award in the Computer Science and Engineering Category of the Board of IT Education Standards in 2009, Ramanujan Lecture presenter award of the Institution of Engineers India (IEI-India) in 2015.