Handling Heterogeneous Data in Knowledge Graphs: A Survey

Sushmita Singh* and Manvi Siwach

Department of Computer Engineering, J.C. BOSE University of Science and Technology, YMCA, Faridabad, Haryana, India
E-mail: sushmi278@gmail.com
**Corresponding Author*

Received 11 August 2021; Accepted 21 February 2022;
Publication 15 April 2022

Abstract

In this era of information where everything is digital, data tends to be ubiquitous. Data Analytics is a term that covers all the areas that deal with the logical analysis of raw data Graph analytics is one of the emerging domains of data analytics that represents and analyses data in the form of knowledge graphs. Knowledge graphs play a vital role in analysing and processing data in order to make decisions. In knowledge graphs the data is stored in the form of entities, relationships between the entities and the attributes of entities as well as attributes of relationships. Construction of knowledge graph and its analytics face multiple challenges like data redundancy, heterogeneity of data, missing data, dynamic nature of real-world data etc. This paper focuses on the issue related to heterogeneity of data while constructing a knowledge graph, and it provides a systematic literature review over construction of knowledge graphs from heterogeneous data sources. This review compiles state-of-the-art knowledge fusion techniques. To conduct this systematic literature review, an exhaustive approach has been adopted to identify various procedures and

Journal of Web Engineering, Vol. 21_4, 1145–1186.
doi: 10.13052/jwe1540-9589.2147

algorithms included and adapted by different research works for knowledge graph construction.

Keywords: Knowledge graph, knowledge fusion, heterogeneous data, entity linking, entity extraction, entity alignment, ontology, knowledge base.

1 Introduction

Data exists in three forms: structured, unstructured and semi-structured. Structured data refers to the organised form of data i.e., when the data is stored in a manner corresponding to a particular schema or data model. Unstructured data refers to free text available in the form of documents, websites, online forms, etc., it has no conformed structure or data model. Semi-structured data refers to data which is not organised in some table or relation but has some features and flexibility to get organised; for example, JSON objects or XML. There exists multiple database software that are designed to handle structured data but most of the data that is generated is either unstructured or semi-structured.

In order to process the other two types of data i.e., unstructured and semi-structured, there are two important requisites. The first requirement is that the data representation method should be able to derive knowledge out of the data collected and stored. Second, it should be able to employ data storage and processing methods that are not schema-bounded i.e., they are not limited by the structural organisation of data. As knowledge graphs possess both of the qualities mentioned above, therefore, they were introduced in the world of data. It is an ideal tool to process and analyse the data and information and then derive the logic that is either explicit or implicit in that data. A knowledge graph is used to answer the 'how' of an event. Thus, it can optimize and enhance the ability of humans to develop explanations correlational or causational explanations of an event. Knowledge graphs are a data analytics tool (graph analytics to be more specific) which is employed to represent and analyse data. Due to knowledge graph capability of retrieving implicit knowledge it can be observed as an integration of a knowledge base (e.g., ontology) with a reasoning engine [Ehrlinger et al.,1]. Knowledge graph is a preferable data analytics tool for heavily linked datasets. It can be precisely defined as follows –

Definition:
A knowledge graph is a network of data entities, which is used to represent and process the data in a way that makes the data analysis easier and more efficient.

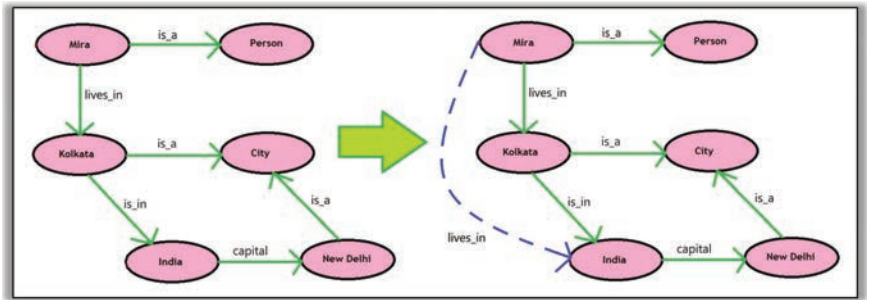

Figure 1 Knowledge graph and Implied knowledge (dotted link) from the knowledge graph.

Now, the data consists of three things: real world objects or concepts; relationships; and properties of these objects. The data in a knowledge graph can be seen in Figure 1 represented in the form of nodes that represent real world objects or concepts and properties (which itself is a real-world concept); edges portray the relationship between these entities. The knowledge stored in these explicit relationships that exist between entities are analysed to derive implicit knowledge. As demonstrated in Figure 1, the links- "Mira lives in Kolkata" and "Kolkata is in India" are explicit in nature and the link- "Mira lives in India" is implied from the two previously mentioned relationships.

However, these relationships can be categorized into two classes: ones that connect two entities (e.g.- "Mira lives in Kolkata", here 'Mira' and 'Kolkata' both are two unique objects of different types which do not constitute each other's characteristic properties) and the others that connect one entity with one of its attributes (e.g.- "Kolkata is a city" here 'City' is an ontological entity that constitutes the type of the other entity i.e.- 'Kolkata'). In order to construct a knowledge graph, the first step is to collect data. Now, one dataset may not be sufficient to solve a particular problem. Therefore, multiple datasets are analysed and processed to generate a knowledge graph that may be sufficient for problem solving in any domain. These datasets vary in terms of format and context. This leads to the issue of heterogeneous datasets. The core of this paper lies in critically analysing the research works that have exploited and investigated this issue as well as different measures to fix it.

1.1 Motivation

Knowledge base is a way to store and represent complex types of data which is cultivated to devise knowledge. Data analysis requires technologies like knowledge graphs that integrate the concept of reasoning with knowledge

base to strengthen the notion of mining and deriving knowledge from the raw information present in the form of facts.

Numerous literature-works exist that depict the construction and design of knowledge bases or knowledge graphs that have applications in multiple domains, although not many discuss the construction of knowledge graphs from heterogeneous data sources. Hence, this paper tries to put some light on the research that focuses on unification of heterogeneous data into a knowledge graph and the challenges that arise in the process of knowledge graph construction from complex and heterogeneous data.

1.2 Organization of Paper

This paper is organized in following manner. Section 2 depicts the evolution of knowledge representation and graphs temporally as well as conceptually. Then, Section 3 compares different survey papers related to the domain of knowledge graphs. Next, Section 4 describes the structure and process of this review. Section 5 gives a survey of knowledge graph construction techniques. After that, Section 6 gives a brief introduction to knowledge fusion which includes two sub-sections. The first subsection gives analysis of various research papers on knowledge extraction and entity linking. The second subsection gives a short review of research papers on entity alignment. Section 7 concludes this paper with future scope of the research work on the issue of heterogeneous data in knowledge graphs and finally, Section 8 is references.

2 Evolution of Knowledge Representation and Graphs

The history of knowledge representation goes back to the 1950s when semantic nets came into existence. In 1956, Richard Hook Richens implemented semantic nets (knowledge base which represents the semantic links between the objects as well as concepts) for the first time [Singhal et al., 2]. Semantic networks were then adapted by Marvin Minsky in 1974 to create Frame networks in an article titled "A Framework for Representing Knowledge" [Minsky, 3]. These frame networks are like data structures that are used to represent stereotypical situations in the form of a network. Next came the popular Entity relationship model which was introduced by Peter Chen in his paper "*The Entity-Relationship Model – Toward a Unified View of Data*" [Chen, 4] which was published in 1976. ER model is used to represent

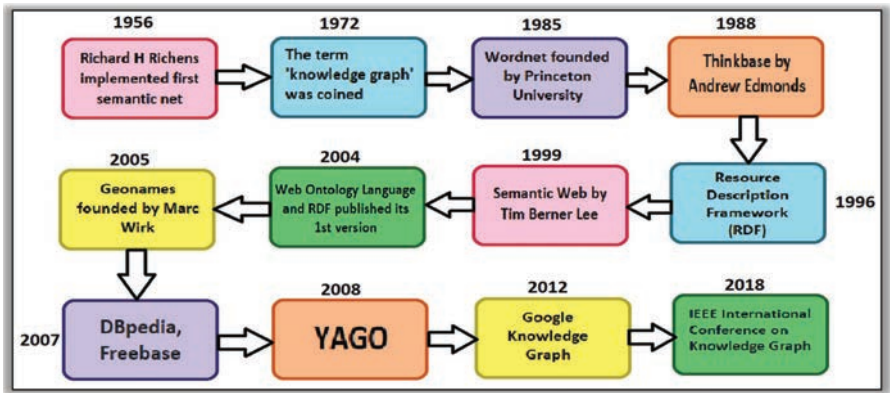

Figure 2 Research timeline of knowledge representation and processing.

structured data. Originally, the term knowledge graph was coined earlier in 1972 [Schneider et al., 5] but it got recognition when Groningen and Twente universities started a joint project named knowledge graphs in 1980's. Then with the coming years this domain of linked knowledge representation was explored.

Figure 2 mentions many of the remarkable years that are related to the history of knowledge representation and knowledge graphs [2–5]. After 2012 when Google upgraded its knowledge representation and processing to knowledge graph then multiple organizations adopted knowledge graphs [Schwartz, 6] for knowledge processing including Amazon, Airbnb, Microsoft, LinkedIn, Uber, etc. In 2019 IEEE combined its annual international conferences on "Big Knowledge" and "Data Mining and Intelligent Computing" into the "International Conference on Knowledge Graph" [7] which is a major step in the research domain of knowledge graphs.

3 Related Surveys

After 2012 when Google adopted knowledge graph for knowledge retrieval [Schwartz 6] and knowledge representation, knowledge graphs secured attention of researchers from multiple domains [Lehman et al., 8]. Therefore, before designing a systematic literature review, it is important to study existing surveys in the domain of knowledge graphs. Different literature surveys offer different aspects like – classifications, comparisons, future directions and taxonomies for knowledge graph representations and learnings. Very

few of the available surveys talk about knowledge graph construction techniques. [Paulheim et al., 9] surveys different knowledge graph refinement approaches. Knowledge graph refinement improves the existing knowledge graphs. The paper divides knowledge graph refinement approaches on the basis of three criteria. First on the basis of goals, the approaches of knowledge graph refinement can be divided into completion and error detection. Second on the basis of targeted information, some approaches target for entity type information and others for relations between entities. Third on the basis of data used, the approaches are either internal or external. Internal methods use knowledge graph itself as the input whereas external methods use knowledge graph as well as other knowledge bases for input. [Zhao et al., 10] surveys the papers on the construction of knowledge graphs and related techniques and tools. This survey mentions heterogeneous and cross-domain knowledge resources as challenges. It also provides a generalized construction procedure. [Goyal et al., 11] gives a taxonomy on knowledge graph embeddings and gives a comprehensive overview of the approaches of knowledge graph embedding as well as applications. Graph embedding refers to the conversion of nodes into low-dimensional vectors in order to store the graph structure information. It also mentions the challenges faced by different graph embedding approaches. A survey by [Xiang et al., 12] focuses on the knowledge graphs created for clinical decision support systems. It also gives a review on the research works classified into two types of graph learning approaches-graph embedding based and path-based. [Ji et al., 13] develops an exhaustive taxonomy on knowledge graph representations and its learning. It also surveys research studies on temporal knowledge graphs and knowledge-aware applications. Survey [Song et al., 14] compiles and summarizes research work on knowledge fusion. The review is conducted over three important factors-open network knowledge fusion and multiple knowledge base knowledge fusion and knowledge evaluation. [Zhao et al., 15] surveys the papers on knowledge fusion from multiple sourced knowledge graphs. It also covers the related areas of open-source knowledge fusion, multi-modal knowledge fusion and information fusion within knowledge graphs.

Table 1 summarizes the contributions and improvements made by different survey papers. However, the sub-domain of knowledge fusion for knowledge graph construction tends to be lesser critiqued by the researchers. This systematic literature review comprehensively compares knowledge graph construction techniques from heterogeneous data hence including research works over knowledge fusion as well.

Table 1 Related surveys

Paper	Year	Time-span	Proposal If Any	Focus Area
Paulheim et al.	2017	1995–2015	An overview and classification of knowledge graph refinement techniques	Knowledge graph refinement approaches
Zhao et al.	2018	2001–2016	a general procedure of knowledge graph construction	Knowledge graph construction techniques
Goyal et al.	2018	1998–2017	Taxonomy on graph embeddings, An open-source Python library, GEM (Graph Embedding Methods)	Knowledge graphs, Graph embeddings and KG applications
Xiang et al.	2019	1994–2018	Future directions in faster query systems for graph databases and	KGs for clinical decisions, Knowledge graph reasoning,
Ji et al.	2020	1999–2019	Taxonomy on Knowledge graph representation and applications; Future directions	Knowledge graph learning, temporal knowledge graph and knowledge-aware applications
Song et al.	2019	1991–2019	Comparison	Knowledge fusion
Zhao et al.	2020	2000–2020	Future directions in knowledge fusion	Knowledge fusion from multiple datasets, multiple models, within knowledge graphs and collaborative reasoning

4 Systematic Literature Review

4.1 Purpose and Process

A literature review is an overview or survey of various research works on a particular topic. There exist different types of literature surveys, out of which systematic literature review confers a predefined and structured review protocol [Uman, 16]. In systematic literature review, a review protocol is followed for selecting quality research studies in relevance to the concerned subject area. The purpose of a systematic literature review is to provide a comprehensive review for a specific problem/argument of a research domain. This literature review aims to acquaint the reader with all the possible

Figure 3 Review process.

techniques being used for entity alignment and entity linking in creating a knowledge graph.

There exist three phases in this systematic literature review process according to which this literature review is designed and developed. First phase is planning which involves identification of research questions and designing the research protocol. Second phase is conducting the review which includes four sub-steps: selecting relevant research, reviewing papers, assessing quality of selected research studies, extracting and synthesizing data. The third and final phase is reporting the review which includes writing the report as well as recommending future research directions.

4.2 Research Issues

During construction of a knowledge graph, extracting and merging data from multiple datasets is a difficult task. These datasets are created in different formats and are heterogeneous in nature. In reference to the problem of heterogeneous data in knowledge graph construction following are some of the research questions that arise –

1. What is the generalized procedure of constructing a knowledge graph?
2. What is the generalized procedure of constructing a knowledge graph from heterogeneous data?
3. What constitutes the fusion of knowledge from heterogeneous data?
4. How to link entities with real-world?

Table 2 Keywords based on research questions

Sr. No.	Research Question	Keywords
1.	What is the generalized procedure of constructing a knowledge graph?	Knowledge graph construction, Triplet generation
2.	What is the generalized procedure of constructing a knowledge graph from heterogeneous data?	Heterogeneous knowledge graphs, heterogeneous data-sources
3.	What constitutes the fusion of knowledge from heterogeneous data?	Knowledge fusion, Data fusion
4.	How to link entities with real-world?	Entity linking, Named-entity recognition, Named-entity disambiguation.
5.	How to connect entities from two different knowledge graphs?	Entity matching, Entity alignment

5. How to connect entities from two different knowledge graphs?

The next step is to identify the keywords based on the research questions mentioned above. The importance of identifying keywords lies in domain specification and indexing of research works. [Uman, 16] Table 2 mentions some important keywords for the developed research questions for knowledge fusion and knowledge graph construction.

4.3 Review Protocol

A review protocol establishes a procedure to be followed while conducting a review. The review protocol acts as a rulebook to direct a systematic literature review. Following is the review protocol adapted for this literature review.

1. Identifying keywords
2. Selection/inclusion criteria – There should be some specific criteria based on which the research material will be selected.
3. Applying inclusion/exclusion based on the inclusion criteria selected.
4. Search strategy – The protocol should provide a list of the databases and other sources used during literature searches to identify potentially relevant studies. This section will also include the search strategy, such as keywords and criteria for the searches.

4.4 Inclusion/Exclusion Criteria

To search for research material for a literature review the most important factor is the relevance of research papers. In order to ensure the relevance

Table 3 Inclusion/Exclusion criteria

Criteria	Inclusion	Exclusion
Type of publication	Review papers, Research articles	News reports, Meta-analysis
Source of publication	Journals, Conferences and Symposiums	Grey literature
Language	English	Other than English
Others	Informative, Domain relevant	Duplicates, Irrelevant

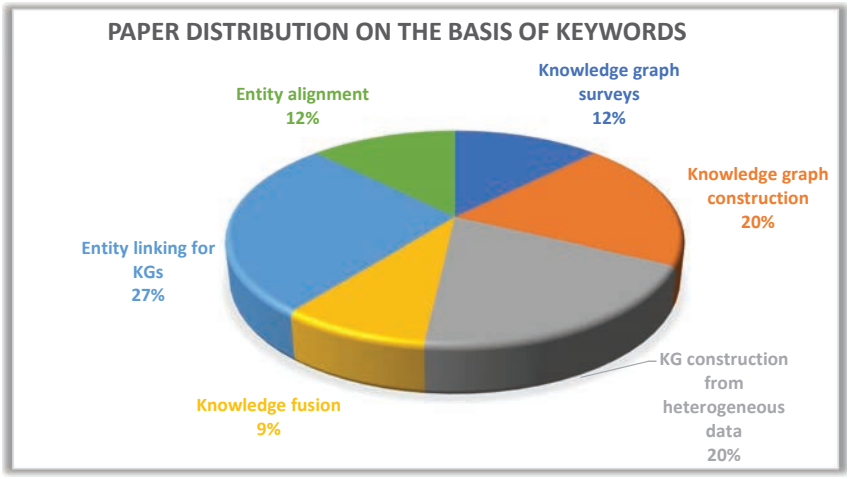

Figure 4 Keyword based distribution of research papers.

factor inclusion/exclusion criteria need to be selected. Inclusion/exclusion criteria helps in setting up the bounds for a focused and systematic research. Table 3 lists down the selected inclusion/exclusion criteria for this literature review.

4.5 Article Selection and Distribution

Keyword based search results in huge number of papers for respective keywords. Then after applying inclusion-exclusion some of these papers 127 were selected. Next, on the basis of topic relevance and quality 38 of them were discarded, leaving 89 for the next step. After studying the abstract 17 out of the 89 were discarded. After a detailed study of papers, an additional 16 of them were discarded. Figure 4 represents the keyword-based distribution of selected papers. Figure 5 represents publisher-based distribution of the selected articles.

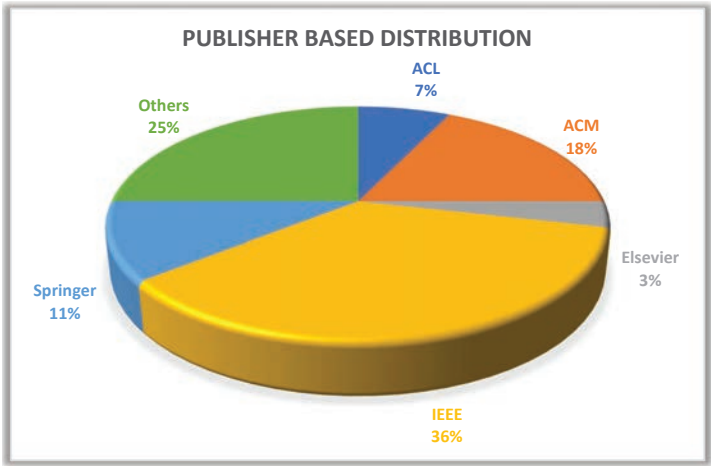

Figure 5 Publisher based distribution of research papers.

5 Review on Knowledge Graph Construction

Knowledge graph construction is a process which involves the conversion of raw data into a representational and conceivable form. Various research papers propose different construction procedures for the knowledge graph construction in detailed manner. This section puts light on such research works that present efficient knowledge graph construction techniques.

5.1 Homogeneous Knowledge Graphs

One of the oldest and well-known knowledge graphs is DBpedia [Auer, 17]. It was constructed to extract the information from Wikipedia and converts it into a web of data that can be accessed and queried semantically. First the content of Wikipedia is converted into RDF(Resource Description Framework) triples, and a multidomain RDF dataset is generated. This dataset is linked with other available open datasets using RDF links.

[Martinez-Rodrigueza, 18] proposes an approach of generating knowledge graphs by using binary relations produced by an Open Information Extraction approach. The approach is based upon Natural Language Processing and Information Extraction. The approach consists of Entity Extraction and Linking (EEL); and Relation Extraction (RE). RE is followed by automatic Semantic Role Labelling (SRL) which defines the semantic roles of extracted relationships. Finally, RDF-triples are created out of the selected components and orderings by putting all the data together.

[Heist, 19] gives a three-phase construction process of knowledge graphs. The main idea proposed in this paper is pattern extraction and pattern fusion. Relationship co-occurrence pattern of the entities is calculated and a set of patterns for each document is generated. It gives an iterative process for construction as well as refinement of knowledge graph.

[Gawriljuk et al., 20] presents an approach to consolidate data from multiple data sources and build a knowledge graph as well as to extend a previously existing knowledge graph. The process consists of five steps. It uses schema mapping and hashing to create a knowledge graph.

[Kertkeidkachorn et al., 21] proposes a text to knowledge graph system called T2KG that constructs knowledge graph from unstructured text. This paper introduces and glorifies the idea of predicate mapping. The architecture of the proposed system consists of five components – (1) Entity Mapping, (2) Coreference Resolution, (3) Triple Extraction, (4) Triple Integration and (5) Predicate Mapping. To deal with heterogeneous vocabularies and to alleviate the sparsity of unstructured text, a hybrid combination of a rule-based approach and a similarity-based approach using the vector-based similarity metric is proposed in this study. [Clancy et al., 22] creates an end-to-end platform for constructing knowledge graphs from unstructured text via the integration of four mature technologies: Apache Solr, Stanford CoreNLP, Apache Spark, and Neo4j. CYPHER query language is also being used to manipulate the graph. According to [Jeyaraj, 23] the construction process of knowledge graph consists of three steps. During the first phase facts are extracted from the free data, then in second phase triples in the form of subject, predicate and the object are created from these extracted facts. Finally, in the third phase knowledge graph is created from this knowledge base. [Das et al., 24] gives a model that constructs dynamic knowledge graph from text. The model explicitly creates a knowledge graph that tracks the changes in the state of text and it also improves the question-answering ability of the text. The machine reading comprehension uses Recurrent Neural Network (RNN) for encoding the text passage and queries. [Li et al., 25] gives a procedure to create medical knowledge graph from electronic medical records. The approach consists of eight phases – (1) data preparation, (2) entity recognition, (3) entity normalization, (4) relation extraction, (5) property calculation, (6) graph cleaning, (7) related-entity ranking, and (8) graph embedding. [Penga et al., 26] suggests that the knowledge can be classified into three hierarchical levels – basic knowledge, deep knowledge and application knowledge. The model consists of a classification agent uses CNN (convolutional neural network). It classifies the knowledge into three

categories and then pass the corresponding knowledge to corresponding knowledge agents. Finally, a hierarchical knowledge graph is constructed from the three obtained graphs. [Zhao et al., 27] proposes a Text-CNN based information extraction model. It also employs feature vectorisation and uses topic information as well as summary information to integrate the data. The construction process involves data vectorisation which uses Word2vec model to extract the feature vector for words. Next step is classification of sentences using Text-CNN.

5.2 Heterogeneous Knowledge Graphs

[Wilcke et al., 28] proposes the idea of considering knowledge graph as a default knowledge representation in order to collect, store and process heterogeneous data. The main concept is that for heterogeneous and multimodal datasets knowledge graphs can prove to be a suitable option. [Liu et al., 29] proposes an architecture for paediatric disease prediction that takes clinical data, text books and expert experiences as the source of data. The concept of hybrid knowledge graph is used here which integrates the data from different dissimilar sources. The architecture also consists of intelligent reasoning module and human computer interaction module. [Wang et al., 30] introduces an approach of designing a knowledge graph that consists of patient centred data (data about a particular patient profile), local knowledge (knowledge that is collected from local hospitals, clinics, etc), global (knowledge that is collected at worldwide level) and different medical instances. Primary focus is to semantically transform the medical records into semantic data and then integrate the PHR (personal health records) system with the medical information management systems of hospitals. [Shi et al., 31] constructs a health knowledge graph which semantically integrates the heterogeneous and vast Textual Medical Knowledge (TMK). The paper proposes a three-layered architecture. It uses textual data mining, pattern extraction and chain inference rules. A document-term matrix is generated for entities and one for each relation, and then classification algorithm is applied to get precise results.

[Kiourtis et al., 32] proposes an architecture which generates a common health language to standardise the medical terms from multiple health information. The architecture consists of four parts – (i) the Data Enrichment component, (ii) the Dataset Classification component, (iii) the Semantic Annotation component, and (iv) the Common Health Language (CHL) component. [Collarana et al., 33] integrates data from multiple heterogeneous

web resources and create a knowledge graph. This is done by implementing a molecule-based integration framework called – Molecule based Integration Technique (MINTE$^+$). The basic data unit of the architecture of MINTE$^+$ is an RDF-molecule. An RDF (Resource Description Framework) molecule is a subgraph which is created by lossless decomposition of a knowledge graph. It basically uses a semantic similarity function to identify equivalent RDF molecules. MapSDI [Jozashoori et al., 34] framework uses a data integration system (that includes the data sources, ontology and mapping rules) as an input, and step wise transforms it into a knowledge graph. Main step is RDFization (transformation of data into RDF) of data which uses transformation rules to convert the pre-processed data into RDF-triple maps. [Zhang et al., 35] highlights the role of neural networks in graph representation learning for heterogeneous knowledge graphs. It proposes a heterogeneous graph neural network called HetGNN which has four components which use the node embeddings of neighbouring nodes. [Vidal et al., 36] formulates a computational framework for the semantic integration of heterogeneous data into a knowledge graph. The framework consists of four main components – (1) Knowledge extraction; (2) Semantic data integration; (3) Exploration and traversal; (4) Knowledge discovery. [Miaol et al., 37] constructs a traditional Chinese medicine prescription knowledge graph from different data sources for diseases, symptoms, medicines. This paper first constructs domain ontology construction to create a knowledge graph. [Zhang et al., 38] proposes an approach to construct a domain-specific multi-modal knowledge graph which includes not only text related content but images as well. It uses DNN (deep neural network) for classification and labelling.

Table 4 summarizes the research works and their contribution towards the construction of knowledge graphs. As some of them are constructing knowledge graph from homogeneous data and some of the techniques process heterogeneous data for constructing knowledge graphs, hence, there doesn't exist any common grounds for the comparison of these knowledge graph construction processes. However, Table 5 lists the essential steps and corresponding techniques exploited in research papers for creating a knowledge graph.

After studying the previously mentioned research works on knowledge graph construction a sequential pipeline can be devised for constructing a knowledge graph. Figure 6 illustrates various steps involved in creating a knowledge graph and mentions different methods and techniques used for each step.

Table 4 Features used in various KG construction methods

Sr. No.	Paper	Year	Handles Heterogeneous Data	Features
1.	DBpedia: A Nucleus for a Web of Open Data [17]	2007	No	• RDF Knowledge graph for Wikipedia • Linked with other datasets of semantic web crawlers
2.	OpenIE-based approach for Knowledge Graph construction from text [18]	2008	No	• Open Information extraction approach • Data pre-processing Semantic labelling
3.	Towards Knowledge Graph Construction from Entity Co-occurrence [19]	2012	No	• Pattern extraction based on nature of relationships (implicit or explicit) • Iterative process to construct and refine knowledge graph. • Cross document working
4.	A Scalable Approach to Incrementally Building Knowledge Graphs [20]	2016	No	• Incremental process for knowledge graph refinement • Uses hashing algorithms to handle scalable data
5.	T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text [21]	2017	No	• Entity mapping from unstructured data to RDF triples • Rule based approach for triple extraction and predicate mapping
6.	Knowledge Graph Construction from Unstructured Text with Applications to Fact Verification and Beyond [22]	2019	No	• Document extraction • Entity enrichment using other datasets like wikipedia

(*Continued*)

Table 4 Continued

Sr. No.	Paper	Year	Handles Heterogeneous Data	Features
7.	Conceptualizing the Knowledge Graph Construction Pipeline [23]	2019	No	• Defined outline of the process
8.	Building Dynamic Knowledge Graphs from Text using Machine Reading Comprehension [24]	2019	No	• Builds Dynamic graphs • Incremental process • Uses Recurrent Neural Networks to track the changes in state of the graph
9.	Real-world data medical knowledge graph: construction and applications [25]	2020	No	• Data pre-processing • Entity ranking based on relevance and connectivity • Graph cleaning and embedding
10.	Construction of hierarchical knowledge graph based on deep learning [26]	2020	No	• Classifies knowledge into three hierarchies • Uses CNN to classify the knowledge into three divisions
11.	Research on Information Extraction of Technical Documents and Construction of Domain Knowledge Graph [27]	2020	No	• Uses word2Vec model for feature vectorization • Text-CNN based information extraction • Topic selection
	Heterogeneous knowledge graphs			
12.	The Knowledge Graph as the Default Data Model for Learning of Heterogeneous Knowledge [28]	2017	Yes	• Position paper in favour of "knowledge graph as a solution for representing, managing and processing heterogeneous data".

(*Continued*)

Table 4 Continued

Sr. No.	Paper	Year	Handles Heterogeneous Data	Features
13.	A Hybrid Knowledge Graph Based Paediatric Disease Prediction System [29]	2017	Yes	• Uses naïve Bayes classifier • Hybrid knowledge graph for diagnosing disease • Reasoning module • Human-computer interaction module
14.	Design and Implementation of Personal Health Record Systems based on Knowledge [30]	2018	Yes	• Focuses on semantic integration of PHR systems with other hospital information systems
15.	Semantic Health Knowledge Graph: Semantic Integration of Heterogeneous Medical Knowledge and Services [31]	2017	Yes	• Knowledge organization as first phase • Textual data mining to extract patterns • Inference rules for query processing
16.	Aggregating Heterogeneous Health Data Through an Ontological Common Health Language [32]	2017	Yes	• Data cleaning and classification using SVM • Generates a common health language for different ontologies
17.	Synthesizing Knowledge Graphs from Web Sources with the MINTE+ Framework [33]	2018	Yes	• Creates RDF molecules (basic sub graphs) • Uses semantic similarity functions • Merging identical RDF molecules
18.	MapSDI: A Scaled-up Semantic Data Integration Framework for Knowledge Graph Creation [34]	2019	Yes	• Semantic integration • Data pre-processing • Duplication removal • Use of relational algebra

(*Continued*)

Table 4 Continued

Sr. No.	Paper	Year	Handles Heterogeneous Data	Features
19.	Heterogeneous Graph Neural Network [35]	2019	Yes	• Samples each neighbours of each node • Uses bi-LSTM to encode the heterogeneity of neighbour nodes • Categorizes the nodes based on the embeddings and features
20.	Semantic Data Integration Techniques for Transforming Big Biomedical Data into Actionable Knowledge [36]	2019	Yes	• Triplet extraction • Association of triplets with description • Semantic data integration with ontologies and vocabularies
21.	Construction of Semantic-based Traditional Chinese Medicine Prescription Knowledge Graph [37]	2019	Yes	• Relationship acquisition using Protégé • Entity alignment • SPARQL for queries
22.	From vision to content: Construction of Domain-specific Multi-modal Knowledge Graph [38]	2020	Yes	• Images and text • Deep neural network for image classification • Machine learning for semantic labelling

After having a closer and critical look at the research works on knowledge graph construction from multiple data sources either homogeneous or heterogeneous, it is important to realise the need for knowledge fusion. Hence, the next section not only explains knowledge fusion, but also investigates and analyses various State-of-the-art approaches of knowledge fusion.

6 Knowledge Fusion

The primary idea behind construction of a knowledge base or a knowledge graph is to accumulate and assemble all the machine-readable data related to the domain of the knowledge graph being created. Therefore, information

Table 5 Research paper distribution based on essential features

Features Identified	Techniques	Bibliography Reference Papers
Knowledge learning	Entity embeddings, Attribute embeddings, Bag of words	[6, 8–10, 21, 29, 32, 33, 53]
Classification	LSTM (Long Short Term Memory Network), SVM (Support Vector Machine), CNN (Convolutional Neural Network), DNN (Deep Neural Network), Naïve-bayes classifier, Random walk with restart (RWR)	[23, 25, 26, 29, 47, 50, 51, 54]
Semantic learning	Semantic similarity, semantic integration, semantic labelling	[4, 8–10, 13, 14, 22, 24–27, 29, 31, 48, 50, 54]

extracted from multiple datasets need to be combined and stored in a single repository which might include conflict resolution and data reconciliation as well [Dong et al., 39]. This is called data fusion which mainly aims to solve data conflicts and tries to find true value of data items. Data fusion refers to the integration of more than one data source to get more accurate and consistent information [Zhao et al., 15]. In this the aim is to find out the triplets (subject, predicate, object) from different datasets that refer to the same data item and same data value. Knowledge fusion on the other hand has the target to associate and reconcile concepts related to same data entity from multiple data sources [Zhao et al., 15]. Data fusion and knowledge fusion both are the tasks of natural language processing. Knowledge fusion means to integrate the information extracted from different data sources and find out the degree of correctness of the extracted fused data by solving or removing the conflicts, redundancy and ambiguities. It is the process of enhancing this data fusion process and adding another dimension to data fusion.

One way to look at knowledge fusion is that it can be divided into two categories on the basis of language:

(a) Mono-lingual entity linking/alignment

In mono-lingual entity linking/alignment the text mentions and entities are from different knowledge bases but both these knowledge bases are in same language.

(b) Cross-lingual entity linking/alignment

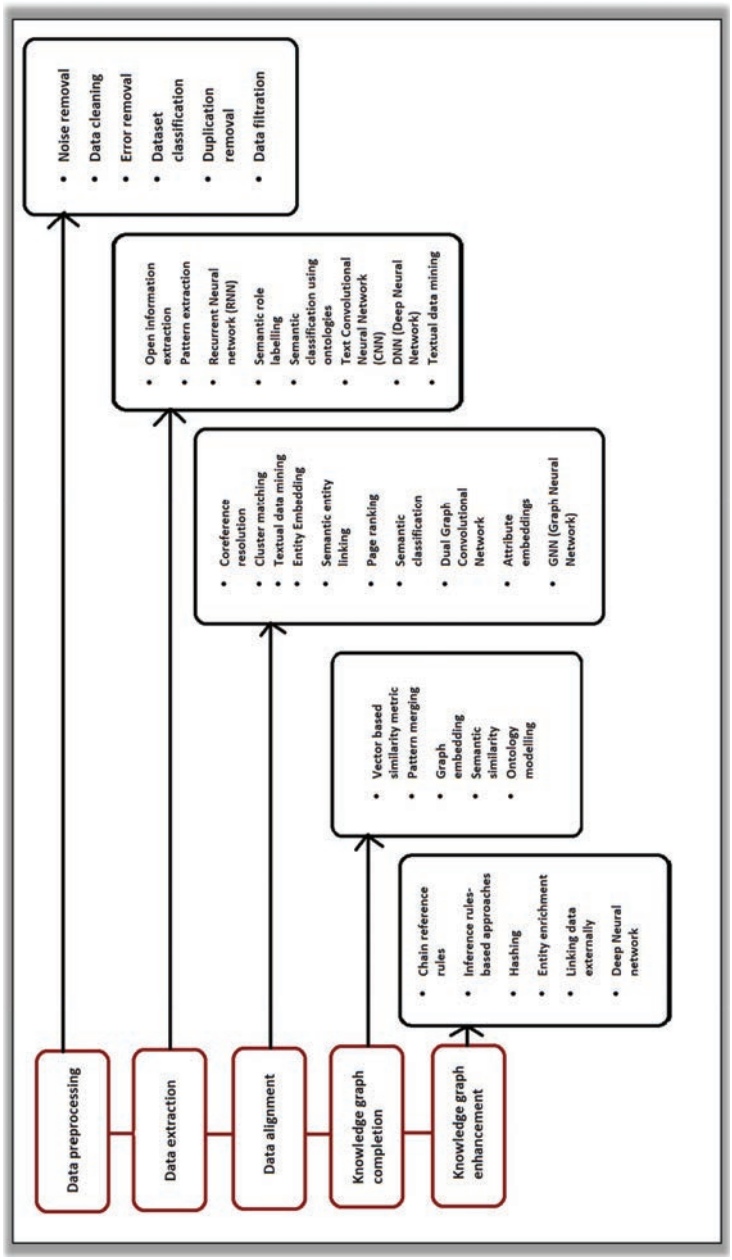

Figure 6 Pipeline for Knowledge graph construction.

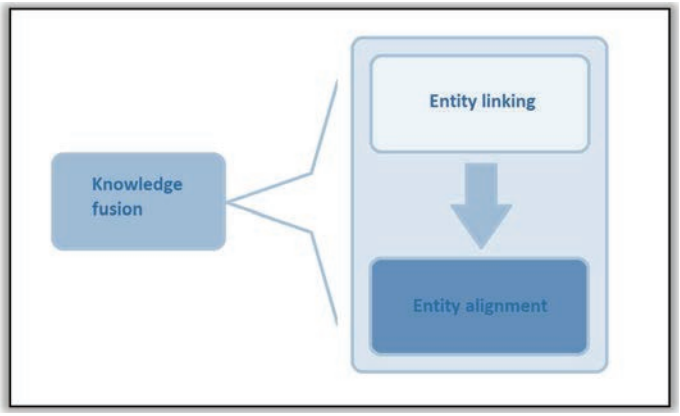

Figure 7 Sequential steps in knowledge fusion.

In cross-lingual entity linking/alignment the text mentions and entities are from different knowledge base and both these knowledge bases are also in different languages.

Knowledge fusion is known by other names like entity alignment, entity resolution, ontology matching [Zhao et al., 15]. Some researches however, treat knowledge fusion and ontology/entity alignment different phases of the process. Knowledge fusion can be visualized as a sequence of knowledge extraction and entity alignment. [Dong et al., 39] suggests that knowledge extraction includes three main steps. The first step is triplet identification, second step is entity linkage and third step is predicate linkage. At the same time paper [Dong et al., 40] infers that there are mainly two phases in knowledge extraction first step is entity linkage which basically means to identify the entities in real world that match with the entities in the triplets. The second step is relation extraction which refers to schema extraction and alignment. Figure 7 illustrates the sequence of steps in knowledge fusion. The following subsections describe entity linking and entity alignment in detail.

6.1 Entity Linking and Knowledge Extraction

Entity linking is a natural language processing (NLP) task under the sub-category of knowledge extraction, whose target is to identify the data entities in a corpus of the dataset (that is being processed) that refers to the real-world entities. It discovers the links that exists between entities in the dataset being analysed and entities from some universal knowledge base.

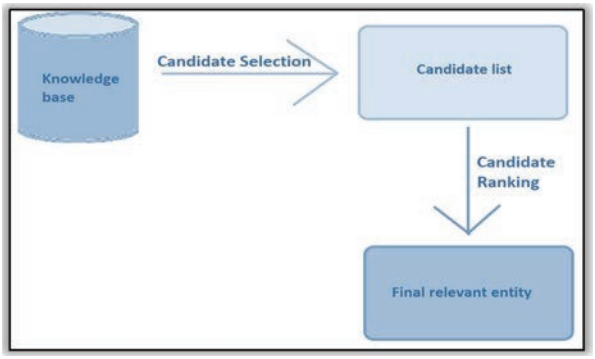

Figure 8 Entity linking.

6.1.1 Process

The process of entity linking can be devised from the definition itself. The generalized procedure of entity linking consists of two main parts, the first part is to select the eligible candidate entities from the universal knowledge base. The second step ranks these selected candidates according to their relevance using some criteria. Finally, based on the ranking the appropriate entity is selected corresponding to the textual mention to be linked. Figure 8 depicts the process of entity linking.

Nagaki et al. suggests that entity linking process can be typically divided into three steps – candidate selection, candidate evaluation and linking decision. Candidate selection tends to be the most important step in entity linking as it can save time. Therefore, next part suggests some candidate selection techniques to make the entity linking process more efficient.

6.1.2 Candidate selection

There exist various approaches for candidate selection. [Song et al., 41] proposes two algorithms for selecting the candidates in the process of entity linking. The first algorithm is called HistSim, which computes the historical similarity by applying a cosine similarity function over the instances and their contexts of all the pairs. Context herein refers to all the paths from one instance to every other instance in the RDF graph. Then selects the candidate pair with history similarity weights greater than a defined threshold value which is based upon a sigmoid function. The second algorithm proposed by [Song et al., 41] is called DisNGram, which selects a candidate selection key to identify maximum disambiguating candidates in a domain independent and unsupervised learning method and then these ontology instances are

arranged to create an index, which is finally looked up for discovering similar instances. These lookup results are again refined to get a smaller set of candidate instance pairs, and finally a character level similarity function is used to determine the candidates. This algorithm also suggests if one attribute is not enough to cover discriminability then combining two attributes as candidate selection key might be helpful. [Nagaki et al., 42] proposes a candidate pruning method which is based on recency of the candidates. Recency of a candidate refers to the strength of dynamic association of a text mention with an entity at a particular time. To calculate the score for entity recency a time window frame is used as a threshold perimeter. Recency based pruning is efficient as it reduces the processing time without decreasing the accuracy. Another approach suggested by [Siedlaczek et al., 43] is Bag-Of-Words for candidate selection and its optimisation for Instance retrieval systems. Different approaches recommend different criteria for selecting the candidate in order to increase the efficiency of entity linking.

6.1.3 Approaches for entity linking

As discussed previously, different criteria are used to select the candidates. Based on the candidate selection criteria as well as candidate ranking criteria there exist different approaches for linking the entities.

[Sil et al., 44] computes granular similarity between the textual mentions from a particular document and the entities in a larger knowledge base, based on the coherence between the entities. This paper proposes a neural network model that computes fine-grained similarity and dissimilarities between the text mentions and entities from more than one perspective and context, by using convolutional neural network, Long Short Term Memory network (LSTM) and neural tensor networks as well. The model comprises of two phases – first phase uses a fast match search algorithm to select candidate entities The second phase ranks the candidates, by calculating a matching score which is computed from some contextual clues and using multi-lingual text embeddings. [Radhakrishnan et al., 45] highlights the concept of density in a knowledge graph and its probable contribution in entity linking. This paper proposes Entity Linking using DENsified knowledge graphs (ELDEN). It is an entity linking system that densifies the entities and links the entities using the embeddings of increased entities and relationships. The knowledge graph is densified by identifying pseudo-entities from the textual and web corpus. Then the next task is to discover relations between these pseudo-entities and the original entities using PMI (Pointwise Mutual Information) measure. Now embeddings are generated for all the entities i.e., both the

original ones and the pseudo-ones which are learned to get the coherence as well as the contextual compatibility between the text mention and the entity. [Zhang et al., 46] divides the entities into three categories and proposes a unified framework – LinKG that has three different neural network-based mechanisms to handle the three categories of entities. First category is word sequence-based entities (e.g., venues), second category is large scale entities (e.g., papers) and the third category belongs to the entities with ambiguity (e.g., authors). The first linking module consists of entity name matching and sequence encoding which is done using long-short term memory networks. For the locality sensitive hashing and convolutional neural networks are used. The third module employs heterogeneous graph attention network and results from the first two modules in order to link the ambiguous entities. [Yin et al., 47] uses BERT model to remove the ambiguity that exists among the words in a particular text and then links these words with real world entities. The input to BERT model consists of two sentences (strings) – one constitutes the text mentions and the context of these text mentions; the other sentence constitutes the entities. The target is to get the semantic relation between the first sentence and second sentence and get the best link possible. In order to do so the first step is to generate entity candidate list and secondly to disambiguate the entities by calculating a score function between the entity and the text mention (first sentence).

[Pershina et al., 48] suggests the personalisation of page rank algorithm for entity linking. It calculates coherence and initial similarity between the entities and combines both of them for ranking score. [Chen et al., 49] proposes a bilinear model – Bilinear joint learning model (BJLM) to learn the entities and words in vector space together by learning their embeddings. [Yamada et al., 50] proposes an extended skip-gram model to calculate word similarity. It suggests to use vector context and their similarity to calculate the coherence between the entities. Both word similarity and context similarity are used for entity linking. [Yamada et al., 51] proposes Neural Text-Entity Encoder (NTEE) which predicts relevant entities for every textual mention in the knowledge base. It is based on neural networks. [Park et al., 52] proposes a two phased named entity recognition approach which uses lexical knowledge. The first phase is term recognition which has three sub-steps – pre-processing (extracts domain-salient terms, morphological patterns), boundary detection and post-processing (collocations are extracted i.e., if any out of bound word collocates with inner terms). The second phase is semantic classification. This phase utilises the knowledge and features extracted in first phase and applies semantic classification on boundary regions identified in previous step. [Liu

et al., 53] suggests the use of greedy search and random walk for entity linking. In this paper first an entity graph is created by connecting the linked entities. Next greedy approach and a modified Monte Carlo random walk is applied to this graph for calculating page rank values for nodes.

Since text mining has experienced an increase in interest, diversified entity linking approaches have been developed to obtain efficiency and accuracy. Researchers working on entity linking which is also known as Named-entity recognition, have probed into multiple domains of machine learning and reasoning like supervised and unsupervised learning, deep learning, neural networks, etc. Below is the report of analysis of some important research works in entity linking. Table 6 summarizes different techniques used for entity linking.

After critical review, Table 7 compares research works based on their performance accuracy over the ConLL dataset.

ConLL is a conference organised by ACL for Natural Language Learning which is conducted solely for language independent named entity recognition. Figure 9 shows a sample of ConLL dataset. The format of ConLL dataset consists of four columns – word, part-of-speech tag, syntactic chunk tag and named entity tag.

Figure 10 Depicts the comparison of macro accuracy and micro accuracy combined of various papers.

6.2 Entity Alignment

Entity alignment refers to the identification of entity pairs which has entities from two different knowledge graphs, that cite the same real-world object or concept. Entity alignment is also known by the name entity matching.

6.2.1 Process

The process of entity alignment is similar to the process for linking entities. First the candidates are selected and then a ranking score is calculated based on some criteria. After calculating the score and rank for all the candidate entities, the best suited entity is selected for alignment. Now the difference between entity linking and entity alignment lies in the fact that in entity linking one universal knowledge base is used to link the entities with real world, whereas in entity alignment the two knowledge bases are being fused to get a new one. The process of entity alignment has four steps: candidate selection, aligning criteria selection, entity ranking based on the criteria selected in second step and select the entity for alignment. Figure 11 depicts a flow chart for the process of entity alignment.

Table 6 Summary of papers on entity linking

Sr. No.	Paper	Year	Data Analytics Techniques Used	Feature
1.	Linking Heterogeneous Data in the Semantic Web Using Scalable and Domain-Independent Candidate Selection	2016	Cosine similarity function, N-grams from Natural Language Processing	Candidate selection based on history similarity; the other is based on disambiguating candidate entities
2.	Neural Cross-Lingual Entity Linking	2018	CNN, LSTM, Neural Tensor flow networks	Candidate selection using similarities and dissimilarities
3.	ELDEN: Improved Entity Linking using Densified Knowledge Graphs	2018	Entity embeddings from Natural Language Processing	Identifies sparsely connected entities, densifies the KG
4.	OAG: Toward Linking Large-scale Heterogeneous Entity Graphs	2019	Long-short term memory network, CNN, Heterogeneous graph attention network	Divides the entities into three types, uses different Data analysis networks to process these entities
5.	Deep Entity Linking via Eliminating Semantic Ambiguity with BERT	2019	Bidirectional Encoder Representations from Transformers	Defines sematic relationships between entities and textual mentions based on the occurrence of mentions
6.	Personalized Page Rank for Named Entity Disambiguation	2015	Coherence, semantic similarity, Page ranking	Combines local and global factors for named entity disambiguation
7.	Bilinear Joint Learning of Word and Entity Embeddings for Entity Linking	2018	Ranking, bilinear-model	Learns entity embeddings and word embeddings and simulate interactions between them
8.	Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation	2016	Skip-gram similarity, Coherence, vector space model	Words and entities are mapped together in vector space
9.	Learning Distributed Representations of Texts and Entities from Knowledge Base	2017	Neural networks	Jointly learns text and knowledge base entities to predict relevant entities
10.	ME-Based Biomedical Named Entity Recognition Using Lexical Knowledge	2006	Pattern analysis, Lexical knowledge	Extracts morphological patterns and applies semantic classification
11.	DBpedia-Based Entity Linking via Greedy Search and Adjusted Monte Carlo Random Walk	2017	Greedy search, monte carlo walk	Entity linking and page rank algorithm to connect the entities

Table 7 Comparison of accuracy percentage of various entity linking models over ConLL

Sr. No.	Paper	Technique Used	Micro Accuracy	Macro Accuracy
1	Pershina et al. [37]	Personalised page ranking	91.77	89.89
2	Yamada et al. [39]	Skip-gram similarity, vector space model	93.1	92.6
3	Luo et al. (PBRTA) [34]	Cosine similarity, NLP	94.7	94.3
4	Chen et al. (PBRTB) [38]	Bilinear joint learning model	93.8	93.5
5	Yamada et. Al. (NTEE) [40]	Neural networks	94.7	94.3
6	X. Yin et al. [31]	Bidirectional Encoder Representations from Transformers	95.04	94.82
7	Priya. R et al. (Elden) [33]	Entity embeddings	93	93.7

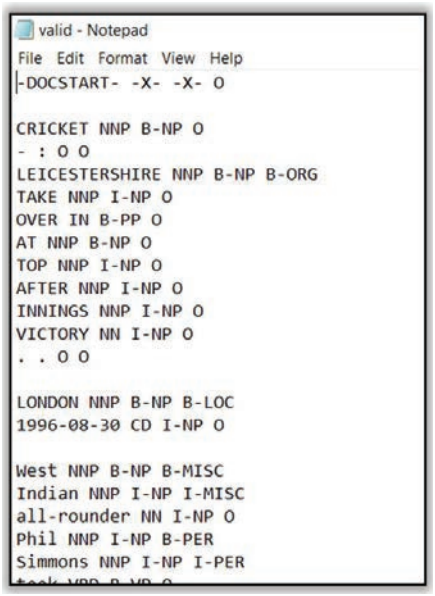

Figure 9 Sample snippet of ConLL dataset [54].

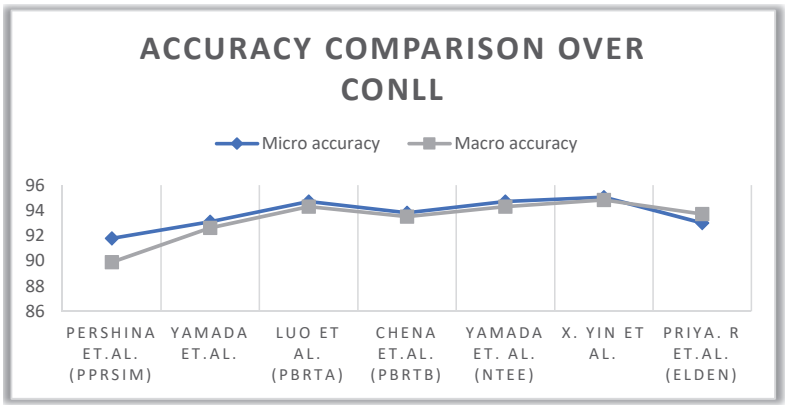

Figure 10 Performance comparison of entity linking techniques over ConLL dataset.

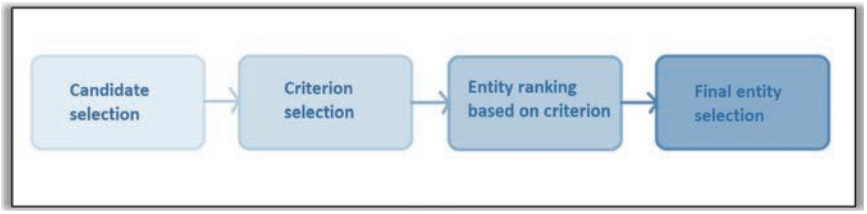

Figure 11 Entity alignment.

Consider two knowledge bases 'KB_1' and 'KB_2'. Now an entity 'e_1' belongs to knowledge base 'KB_1'. Now to align entity e_1 from KB_1 with KB_2, candidate entities are selected from KB_2 and a set of candidate entities is obtained. Next, obtained set is ranked according the alignment criterion. Highest ranked entity, let it be 'e_2', is selected from the set of candidate entities. The desired output is the entity pair – $\{e_1, e_2\}$, where e_2 is the most relevant entity from KB_2 for e_1.

6.2.2 Criteria for Alignment

Different researchers suggest different criteria for entity alignment. Based on the conducted review, following are various criteria based on which entities from two different knowledge bases can be aligned together:

- Structure of entity [Zeng et al., 55]
- Literal meaning of entity [Costa et al., 56]
- Semantic context of entity [Yan et al., 61]
- Degree of the entities [Zeng et al., 55]

- Attribute of entities [Costa et al., 56],
- Frequency of occurrence of entities [Zeng et al., 55]
- Relationships and Predicates [Wu et al., 58] [Zhu et al., 59]
- Attribute attention [Costa et al., 56],
- Neighbourhood of entities Zhu et al., 59

6.2.3 Approaches

[Zeng et al., 55] advances with an approach that uses the concept of degrees in entity structure learning phase of the procedure of entity alignment. The architecture is divided into three phases – pre alignment phase, alignment phase and post alignment phase. In the pre alignment phase there are two modules – name representation learning (textual similarity as well as semantic similarity of entities), and structural representation learning. In the alignment phase degree information is included to extract optimal information for feature modelling. With the help of degree related information, the entities are divided into two categories long-tail and short-tail entities. For long tail entities name representation is given more importance and for the second category of entities structural representation is given more importance. A co-attention feature similarity calculating mechanism is used to do this by creating a similarity matrix between the features of both the entities. And also, weights are assigned in this matrix for the features to determine the corresponding relevance.

[Costa et al., 56] suggests the concept of entity alignment by enriching the entity embeddings with the literal information related to the entity itself. Literal information refers to the non-figurative context or sense of a word or a set of words. This approach considers literals from attributes as well as literals from triplets. The proposed approach exploits linguistic frames and their inclusion in the entity alignment process; it uses a FrameBase schema to map external knowledge bases with the entities and also to integrate the information derived from literals. Linguistic frames depict the meaning of a sentence as a scenario with multiple participants and their semantic roles. Thus, lexical patterns are discovered between the triplets which captures varied correspondence and corelations between literals.

[Trisedya et al., 57] focuses on the task of entity alignment and it also identifies two major challenges in entity alignment – first the inductive knowledge is ignored i.e. the knowledge is derived from relations only and not the other attributes; second the existing methods may fail for entities with sparse connections. This paper proposes the usage of GNN – Graph Neural Network in entity alignment and introduces a Collective Graph neural network for

multi-type entity Alignment, called CGMuAlign. It collects positive as well as negative evidences from neighbourhood in a recursive manner to derive inductive knowledge and attain precision in the identified entity sets (from heterogeneous knowledge graphs) for a single real-world entity.

[Wu et al., 58] proposes an entity alignment model based upon attribute embeddings. The model consists of three submodules – predicate alignment module, embedding learning module and entity alignment module. First module tries to identify partially similar predicates in order to get a unified vector space for relationship embeddings. The core part lies in the second module that involves learning the entity embeddings trough two ways – structure embedding that utilizes the relationship triplets (subject, predicate, object) and; attribute embeddings that utilizes the attribute triplets (subject, predicate, attribute). Finally, in the third module a similarity score is calculated for every pair of entities and pairs with a minimum threshold similarity score are selected. An attribute enrichment approach is also employed in the end to get more robust attribute embeddings.

[Zhu et al., 59] proposes Relation-aware Dual Graph Convolutional Network (RDGCN) to discover the entity pairs (created from different knowledge graphs) that refer to same real-world entity. RDGCN first creates a primal graph by applying simple UNION operation between the knowledge graphs. In the next stage a dual graph is created as a counterpart of primal graph, which represents connectivity measure of two relationships (or triplets) from different graphs. In a dual graph each vertex represents the triplet (or relationship) of primal graph and a vertex is connected to another vertex if they both have either a common head (subject) or tail (object). Weights are also assigned to the edges in a dual graph, based on the similar heads and tails. Lastly a graph attention mechanism is used to obtain vertex representation in the dual relation graph. These representations are obtained by multiple iterative interactions of dual graph and the primal graph, which have two layers for each interaction i.e. dual attention layer and primal attention layer. Finally, entity pairs are obtained from the distance between these vertex representations.

[Huang et al., 60] proposes a framework for entity alignment which emphasises on the context of the entities for alignment. Therefore, it uses not just the entity embeddings but also the attributes of entities that need to be aligned in order to semantically learn high-level semantics of the entity. The framework has three main modules – Entity Topic Learning (TL) module, Structure Embedding (SE) module, and Context Modelling (CM) module. Topic model is a tool of natural language processing and machine learning

Table 8 Comparison of entity alignment research papers

Sr. No.	Paper	Year	Data Analytics Techniques Used	Feature Criteria
1.	Zeng et al. [55]	2017	Structure embedding (From entity embedding), Textual and Semantic similarity matrix.	Degree of the entity, frequently visited entities
2.	Costa et al. [56]	2018	Linguistic frames (from Natural Language Processing)	Attribute of entities, entity embeddings
3.	Trisedya et al. [57]	2019	Attribute embeddings (as well as entity embedding)	Predicates/relationships, learned embeddings
4.	Wu et al. [58]	2019	Dual Graph Convolutional Network	Relationships between the entities (scored by creating a dual graph of relationships)
5.	Zhu et al. [59]	2020	GNN (Graph Neural Network)	Neighbourhood of entities
6.	Huang et al. [60]	2020	Stochastic gradient descent (from entity embeddings), Semantic aggregation.	Attributes of entities, attribute attention
7.	Yan et al. [61]	2020	Text clustering (from text mining), Topic modelling (from NLP), Multichannel CNN (Convolutional Neural Network).	Context of the entities; Topics of entities

which identifies or discovers a "topic" for a set of words. Next module is structure embedding which determines and learns the structure of entities and relationships by using translation-based embeddings and then scores the triplets to measure the plausibility of the triplets. Third module is Context modelling which uses multichannel Convolutional Neural Network to shape the context of an entity. Context here refers to the neighbouring entities and their relation to the entity that has to be aligned. While ranking the target entities for the source entities, Context and Topic enhanced Entity Alignment(CTEA) considers the combined embeddings and representations as well as topic distribution of the attributes as the base criteria. [Yan et al., 61] introduces Entity Alignment based on Semantic Aggregation (EASA) –

Table 9 Performance comparison of entity alignment models

Paper Name	Knowledge Base	Hits@10
RDGCN [59]	DBP FR-EN	95.72
	DBP ZH-EN	84.55
	DBP JA-EN	89.54
N-GRAM [58]	DBP YAGO	95.59
	DBP GEO	92.71
	DBP LGD	93.21
EASA [61]	DBP YAGO	98.57
	Baidu HUDONG	94.89
DAT [55]	DBP YAGO	98.6
	DBP FR-EN	89.9
CTEA [60]	DBP FR-EN	92.3
	DBP ZH-EN	90.5
	DBP JA-EN	91.4

an entity alignment algorithm which uses semantic aggregation of entities as the base concept for entity alignment. Aggregation of semantics are deduced from attributes of the entities and attribute values; which improvises the process of entity alignment. This paper also introduces one other concept called attribute attention which highlights the discrimination of importance among the multiple attributes of a single entity. Therefore, weights are calculated for different attributes belonging to a particular entity. On the basis of these weights semantic aggregation of entities are computed (as mentioned earlier). In the next phase similarity scores are calculated between the entities of different knowledge graphs. And finally, the entities with highest similarity score are aligned together. This paper uses stochastic gradient descent to optimize the entity embeddings.

Table 8 lists down the data analytics tools and technologies used in entity alignment (throughout the process) and it also states the one or more features of a knowledge graph that are used as the primary criteria to align the entities.

Now, Table 9 analyzes and compares the performance of various entity alignment models on different knowledge bases namely DBPedia (English), DBPedia (French), DBPedia (Japanese), DBPedia (Chinese), DBPedia (German), YAGO, Baidu, Hudong. The metric used to compare the performance is Hits@10. It refers to the fraction of positive hits among the top 10 ranked items. The formula to calculate hits@10 is –

Hits@10= (number of correct entities)/(top 10 ranked candidate entities)

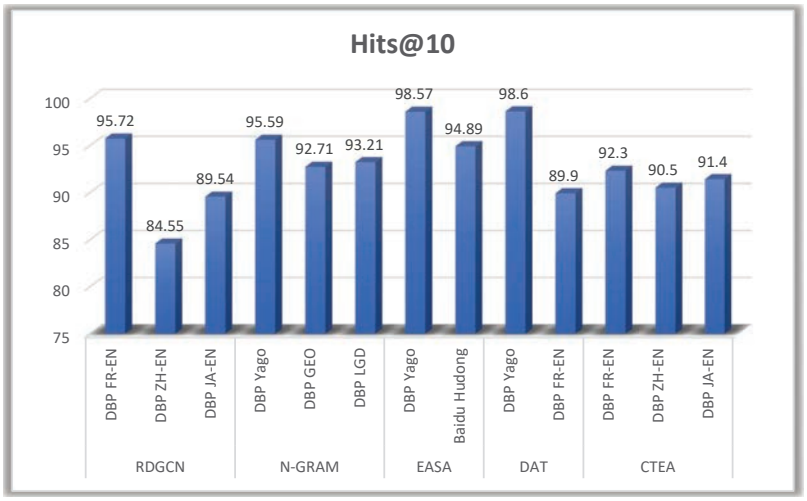

Figure 12 Performance of entity alignment models over different knowledge bases.

Figure 12 depicts the comparison of performance of entity alignment models.

7 Discussions and Conclusion

Data analytics (graph analytics to be more specific) employs knowledge graphs to represent and analyse data effectively. Knowledge graph is useful in retrieving the implicit knowledge from the knowledge base and then deriving inferences and insights using reasoning techniques. It is an efficient representation of heavily linked data. This paper provides a systematic literature review on construction of knowledge graphs. In order to construct a knowledge graph more than one dataset are utilized. The issue arises when these datasets are heterogeneous in nature. Therefore, this paper gives a literature review of significant research works over heterogeneous data and annotates different techniques used to construct knowledge graph from heterogeneous data. Knowledge fusion comes up in the process of handling heterogeneous data. Knowledge fusion consolidates the knowledge from multiple knowledge bases. This paper gives an analytical review of knowledge fusion, which includes entity linking and entity alignment. Entity linking/named entity recognition refers to link the entities extracted from the dataset with the named entities i.e., the real-world objects or concepts. Entity alignment involves entities from two datasets which are linked together

with a real-world object/concept. These two concepts of knowledge fusion constitute important phases of knowledge graph construction. The future holds an enormous number of outlooks and opportunities to explore different other technologies like Cloud computing, Internet of Things, Blockchain [66] etc, merged with knowledge graphs. Knowledge graphs tend to improvise the analysis of data, hence domains that deal with data storage and processing can incorporate knowledge graphs for analytics.

References

[1] Lisa Ehrlinger and Wolfram Wöß. 2016. Towards a Definition of Knowledge Graphs. Semantic-web-journal.

[2] Singhal, Amit (May 16, 2012). "*Introducing the Knowledge Graph: Things, Not Strings*". Google Official Blog. Retrieved September 6, 2014.

[3] Marvin Minsky, 1974. A Framework for Representing Knowledge. Memo no.306 Artificial Intelligence, MIT.

[4] Peter Chen. 1976. The Entity-Relationship Model – Toward a Unified View of Data. ACM Transactions on Database Systems, Volume-1, Issue-1. https://doi.org/10.1145/320434.320440

[5] Edward W. Schneider. 1973. Course Modularization Applied: The Interface System and Its Implications For Sequence Control and Data Analysis. In Association for the Development of Instructional Systems (ADIS), Chicago, Illinois, April 1972.

[6] Schwartz, Barry (December 17, 2014). "Google's Freebase To Close After Migrating To Wikidata: Knowledge Graph Impact?". Search Engine Roundtable. Retrieved December 10, 2017.

[7] International conference of knowledge graphs. World Academy of Science, Engineering and Technology.

[8] Lehmann, Fritz; Rodin, Ervin Y., eds. (1992). Semantic networks in artificial intelligence. International series in modern applied mathematics and computer science. 24. Oxford; New York: Pergamon Press. p. 6. ISBN 978-0080420127. OCLC 26391254.

[9] Heiko Paulheim. 2017. Automatic graph refinement: A survey of approaches and evaluation method. Semantic Web, IOS Press.

[10] Zhanfang Zhao, Sung-Kook Han, In-Mi So. 2018. Architecture of Knowledge Graph Construction Techniques. International Journal of Pure and Applied Mathematics. https://acadpubl.eu/jsi/2018-118-19/articles/19b/24.pdf.

[11] Palash Goyal, Emilio Ferrara. 2018. Graph Embedding Techniques, Applications, and Performance: A Survey. Knowledge-Based Systems. https://www.sciencedirect.com/science/article/abs/pii/S0950705118301540.

[12] Xiayu Xiang, Zhongru Wang, Yan Jia, Binxing Fang. 2019. Knowledge Graph-based Clinical Decision Support System Reasoning: A Survey. IEEE Fourth International Conference on Data Science in Cyberspace (DSC). https://ieeexplore.ieee.org/document/8923457.

[13] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, Philip S. Yu. A Survey on Knowledge Graphs: Representation, Acquisition and Applications. IEEE Transactions on Neural Networks and Learning Systems. https://ieeexplore.ieee.org/document/9416312.

[14] Yichen Song, Aiping Li, Yan Jia, Jiuming Huang, Xiaojuan Zhao. 2019. Knowledge Fusion: Introduction of Concepts and Techniques. IEEE Fourth International Conference on Data Science in Cyberspace (DSC). https://ieeexplore.ieee.org/document/8923715.

[15] Xiaojuan Zhao, Yan Jia, Aiping Li, Rong Jiang, Yichen Song. 2020. Multi-source knowledge fusion: a survey. World Wide Web (2020) 23:2567–2592. https://ieeexplore.ieee.org/document/8923525.

[16] Uman, L.S., 2011. Systematic reviews and meta-analyses. Journal of the Canadian Academy of Child and Adolescent Psychiatry. https://doi.org/10.30701/ijc.v39isuppl_b.856

[17] Soren Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak and Zachary Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. The semantic web, Springer. https://link.springer.com/chapter/10.1007/978-3-540-76298-0_52

[18] Jose L. Martinez-Rodrigueza, Ivan Lopez-Arevaloa, Ana B. Rios-Alvaradob. 2008. OpenIE-based approach for Knowledge Graph construction from text. Expert Systems with Applications, Elseveir https://www.sciencedirect.com/science/article/abs/pii/S0957417418304329.

[19] Nicolas Heist. 2012. Towards Knowledge Graph Construction from Entity Co-occurrence. DBLP, EKAW.

[20] Gleb Gawriljuk, Andreas Harth, Craig A. Knoblock and Pedro Szekely. 2016. A Scalable Approach to Incrementally Building Knowledge Graphs. Springer International Publishing Switzerland. https://link.springer.com/chapter/10.1007%2F978-3-319-43997-6_15.

[21] Natthawut Kertkeidkachorn, Ryutaro Ichise. 2017. T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text

(The AAAI-17 Workshop on Knowledge-Based Techniques for Problem Solving and Reasoning WS-17-12

[22] Ryan Clancy, Ihab F. Ilyas, and Jimmy Lin. 2019. Knowledge Graph Construction from Unstructured Text with Applications to Fact Verification and Beyond. https://aclanthology.org/D19-6607.pdf

[23] Nayantara Jeyaraj. 2019. Conceptualizing the Knowledge Graph Construction Pipeline. Towards Data Science.

[24] Rajarshi Das, Tsendsuren Munkhdalai, Xingdi Yuan, Adam Trischler, Andrew McCallum. 2019. Building Dynamic Knowledge Graphs from Text using Machine Reading Comprehension. ICLR 2019 Conference Blind Submission. https://arxiv.org/abs/1810.05682v1.

[25] Linfeng Li, Peng Wang, Jun Yan, Yao Wang, Simin Li, Jinpeng Jiang, Zhe Sun, Buzhou Tang, Tsung-Hui Chang, Shenghui Wang, Yuting Liu. 2020. Real-world data medical knowledge graph: construction and applications. Artificial Intelligence In Medicin. https://www.sciencedirect.com/science/article/pii/S0933365719309546

[26] Zuquan Penga, Huazhu Songb*, Xiaohan Zhengc, Luotianhao. 2020. Construction of hierarchical knowledge graph based on deep learning. IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). https://ieeexplore.ieee.org/document/9181920.

[27] Huaxuan Zhao, Yueling Pan, And Feng Yang. 2020. Research on Information Extraction of Technical Documents and Construction of Domain Knowledge Graph volume 8, IEEE Access. https://ieeexplore.ieee.org/document/9195862.

[28] Xander Wilcke, Peter Bloem and Victor de Boer. 2017. The Knowledge Graph as the Default Data Model for Learning of Heterogeneous Knowledge. Data Science, vol. 1, no. 1–2, pp. 39–57, IOS Press. https://content.iospress.com/articles/data-science/ds007.

[29] P. Liu, Xiaoqing Wang X. Sun, Xiang Shen, Xu Chen, Yuzhong Sun, Yanjun Pan. 2017. A Hybrid Knowledge Graph Based Paediatric Disease Prediction System. Springer International Publishing AG. https://link.springer.com/chapter/10.1007%2F978-3-319-59858-1_8

[30] Huaqiong Wang, Xiaoyu Miao and Pan Yang. 2018. Design and Implementation of Personal Health Record Systems based on Knowledge. 9th International Conference on Information Technology in Medicine and Education, IEEE computer society. https://ieeexplore.ieee.org/document/8589271.

[31] Longxiang Shi, Shijian Li, Xiaoran Yang, Jiaheng Qi, Gang Pan, and Binbin Zhou. 2017. Semantic Health Knowledge Graph: Semantic

Integration of Heterogeneous Medical Knowledge and Services. BioMed Research International Volume 2017, Article ID 2858423. https://www.hindawi.com/journals/bmri/2017/2858423/.

[32] Athanasios Kiourtis, Argyro Mavrogiorgou, Dimosthenis Kyriazis. 2017. Aggregating Heterogeneous Health Data Through an Ontological Common Health Language. IEEE International Conference on Developments in eSystems Engineering. https://ieeexplore.ieee.org/document/8285817.

[33] Diego Collarana, Mikhail Galkin, Christoph Lange, Simon Scerri, Sören Auer and Maria-Esther Vidal. 2018. Synthesizing Knowledge Graphs from Web Sources with the MINTE+ Framework. Springer Nature Switzerland AG 2018. https://link.springer.com/chapter/10.1007/978-3-030-00668-6_22.

[34] Samaneh Jozashoori, Maria-Esther Vidal. 2019. MapSDI: A Scaled-up Semantic Data Integration Framework for Knowledge Graph Creation. OTM 2019 Conferences, Springer International. https://arxiv.org/abs/1909.01032v1.

[35] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, Nitesh V. Chawla. 2019. Heterogeneous Graph Neural Network. KDD '19, August 4–8, 2019, Anchorage, AK, USA, ACM. https://aclanthology.org/2020.coling-main.29.pdf.

[36] Maria-Esther Vidal, Samaneh Jozashoori, Ahmad Sakor. 2019. Semantic Data Integration Techniques for Transforming Big Biomedical Data into Actionable Knowledge. IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). https://ieeexplore.ieee.org/document/8787394.

[37] Fang Miao1, Huixin Liu1, Yamei Huang1,Chenming Liu2, Xinyi Wu2. 2018. Construction of Semantic-based Traditional Chinese Medicine Prescription Knowledge Graph. IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC 2018). https://ieeexplore.ieee.org/document/8577236.

[38] Xiaoming Zhang, Xiaoling Sun*, Chunjie Xie, And Bing Lun. 2019 From vision to content: Construction of Domain-specific Multi-modal Knowledge Graph. 2933370, IEEE Access. https://ieeexplore.ieee.org/document/8788525.

[39] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Kevin Murphy, Shaohua Sun, Wei Zhang. 2014. From Data Fusion to Knowledge Fusion. Proceedings of the VLDB Endowment, Vol. 7, No. 10 VLDB Endowment 21508097/14/06.

[40] Xin Luna Dong, Divesh Srivastava. 2015. Knowledge Curation and Knowledge Fusion: Challenges, Models, and Applications. SIGMOD'15, May 31–June 4, 2015, Melbourne, Victoria, Australia. ACM. https://dl.acm.org/doi/10.1145/2723372.2731083.

[41] Dezhao Song, Yi Luo and Jeff Heflin. 2016. Linking Heterogeneous Data in the Semantic Web Using Scalable and Domain-Independent Candidate Selection. IEEE Transactions on Knowledge and Data Engineering. https://ieeexplore.ieee.org/document/7562437.

[42] Saki Nagaki, Hiroyuki Kitagawa. 2017. Recency-based Candidate Selection for Efficient Entity Linking. iiWAS, December 4–6, 2017, Salzburg, Austria 2017 Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3151759.3151771.

[43] Michał Siedlaczek, Qi Wang, Yen Yu Cheng, Torsten Suel. 2018. Fast Bag-Of-Words Candidate Selection in Content-Based Instance Retrieval Systems. IEEE International Conference on Big Data (Big Data). https://ieeexplore.ieee.org/document/8621935.

[44] Avirup Sil, Gourab Kundu, Radu Florian and Wael Hamza. 2018. Neural Cross-Lingual Entity Linking. Association for the Advancement of Artificial Intelligence. https://arxiv.org/abs/1712.01813.

[45] Priya Radhakrishnan, Partha Talukdar and Vasudeva Varma. 2018. ELDEN: Improved Entity Linking using Densified Knowledge Graphs. NAACL – 2018 Association for Computational Linguistics. https://aclanthology.org/N18-1167/.

[46] Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xiaotao Gu, Yan Wang, Bin Shao, Rui Li, and Kuansan Wang. 2019. OAG: Toward Linking Large-scale Heterogeneous Entity Graphs. KDD-19, ACM. https://dl.acm.org/doi/10.1145/3292500.3330785.

[47] Xiaoyao Yin, Yangchen Huang, Bin Zhou, Aiping Li, Long Lan, And Yan Jia. 2019. Deep Entity Linking via Eliminating Semantic Ambiguity with BERT. IEEE Access. https://ieeexplore.ieee.org/abstract/document/8911323.

[48] Maria Pershina Yifan He Ralph Grishman) In: Human Language Technologies. 2015. Personalized Page Rank for Named Entity Disambiguation. The 2015 Annual Conference of the North American Chapter of the ACL. https://aclanthology.org/N15-1026.pdf.

[49] Hui Chen, Baogang Wei, Yonghuai Liu, Yiming Li, Jifang Yu, Wenhao Zhu. 2017. Bilinear Joint Learning of Word and Entity Embeddings for Entity Linking. Neurocomputing. https://www.sciencedirect.com/science/article/abs/pii/S0925231217318234.

[50] Ikuya Yamada, Hiroyuki Shindo,Hideaki Takeda, Yoshiyasu Takefuji. 2016. Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation. ACL. https://aclanthology.org/K16-1025.pdf.

[51] Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji. 2017. Transactions of the Association for Computational Linguistics, vol. 5, pp. 397–411. https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00069/43409/.

[52] Kyung-Mi Park, Seon-Ho Kim, Hae-Chang Rim. 2006. ME-Based Biomedical Named Entity Recognition Using Lexical Knowledge. ACM Transactions on Asian Language Information Processing, Vol. 5, No. 1, March 2006, Pages 4–21. https://dl.acm.org/doi/abs/10.1145/1131348.1131350.

[53] Ming Liu, Lei Chen, Bingquan Liu, Guidong Zheng, And Xiaoming Zhang. 2017. DBpedia-Based Entity Linking via Greedy Search and Adjusted Monte Carlo Random Walk. ACM Transactions on Information Systems, Vol. 36, No. 2, Article 16. https://dl.acm.org/doi/10.1145/3086703

[54] Source: https://www.kaggle.com/alaakhaled/conll003-englishversion

[55] Weixin Zeng, Xiang Zhao, Wei Wang, Jiuyang Tang, Zhen Tan. 2017. Degree-Aware Alignment for Entities in Tail. Proceedings of ACM Conference, Washington, DC, USA. https://dl.acm.org/doi/10.1145/3397271.3401161.

[56] Gustavo de Assis Costa, José Maria Parente de Oliveira. 2018. Linguistic Frames as Support for Entity Alignment in Knowledge Graphs. 20th International Conference on Information Integration and Web-based Applications & Services, ACM, New York, NY, USA. https://dl.acm.org/doi/abs/10.1145/3282373.3282415.

[57] Qi Zhu, Hao Wei, Bunyamin Sisman, Da Zheng, Christos Faloutsos, Xin Luna Dong, Jiawei Han. 2020 Collective Multi-type Entity Alignment Between Knowledge Graphs Proceedings of The Web Conference 2020 (WWW'20), April 20–24, 2020, Taipei, Taiwan. ACM, New York, NY, USA. https://dl.acm.org/doi/fullHtml/10.1145/3366423.3380289.

[58] Bayu Distiawan Trisedya and Jianzhong Qi and Rui Zhang. 2019. Entity Alignment between Knowledge Graphs Using Attribute Embeddings. Association for the Advancement of Artificial Intelligence. https://doi.org/10.1609/aaai.v33i01.3301297.

[59] Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan and Dongyan Zhao. 2019. Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs. Proceedings of the Twenty-Eighth

International Joint Conference on Artificial Intelligence. https://doi.org/10.24963/ijcai.2019/733.

[60] Li-an Huang and Xiangfeng Luo. 2020. EASA: Entity Alignment Algorithm Based on Semantic Aggregation and Attribute Attention. IEEE Access. https://ieeexplore.ieee.org/document/8966369.

[61] Zhihuan Yan, Rong Peng, Yaqian Wang, Weidong Li. 2020. CTEA: Context and Topic Enhanced Entity Alignment for knowledge Graphs. Neurocomputing, Elsevier. https://doi.org/10.1016/j.neucom.2020.06.054.

[62] Yichen Song, Aiping Li, Yan Jia, Jiuming Huang, Xiaojuan Zhao. 2019. Knowledge Fusion: Introduction of Concepts and Techniques. IEEE Fourth International Conference on Data Science in Cyberspace. https://ieeexplore.ieee.org/document/8923715.

[63] Aoran Li, Xinmeng Wang, Wenhuan Wang, Anman Zhang and Bohan Li. 2019. A Survey of Relation Extraction of Knowledge Graphs. Springer Nature Switzerland AG. https://link.springer.com/chapter/10.1007%2F978-3-030-33982-1_5.

[64] Isaiah Onando Mulang, Kuldeep Singh, Chaitali Prabhu, Abhishek Nadgeri, Johannes Hoffart, Jens Lehmann. 2020. Evaluating the Impact of Knowledge Graph Context on Entity Disambiguation Models. CIKM '20, October 19–23, 2020, Virtual Event, Ireland ACM. https://dl.acm.org/doi/10.1145/3340531.3412159.

[65] Wan Tao, Qi Zhou, Yuqian Zhao, Aolong Yu. 2020. A Cross-Field Construction Method of Chinese Tourism Knowledge Graph based on Expansion and Adjustment of Entities. IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC 2020). https://ieeexplore.ieee.org/document/9141655.

[66] Shuai Wang, Chenchen Huang, Juanjuan Li, Yong Yuan, Fei-Yue Wang. 2019. Decentralized Construction of Knowledge Graphs for Deep Recommender Systems Based on Blockchain-Powered Smart Contracts IEEE Access volume 7. https://ieeexplore.ieee.org/document/8844724

Biographies

Sushmita Singh is a Junior Research Fellow (JRF scholar), currently pursuing a Ph.D. in Data Analytics at the Department of Computer Engineering, JC Bose University of Science and Technology, YMCA, India. She did her M.Tech. (Information Technology) from the same university in the year 2016, and B.Tech.(Computer Science) from the School of Engineering and Sciences, BPS Women's University, India in the year 2014. She has more than 3 years of experience as an Assistant Professor in colleges and universities. She has published multiple research papers in international journals.

Manvi Siwach, Assistant Professor in department of Computer Applications has completed her Ph.D. in Information retrieval in 2017 from J.C. Bose University of Science & Technology, YMCA, Faridabad. She did her M.Tech.(Computer Engineering) from YMCA University of Science and Technology in year 2008, and B.Tech.(Computer Science) from Kurukshetra University, Kurukshetra in 2005. She has guided more than 15 M.Tech thesis and currently 2 candidates are pursuing their Doctorate under her.

She has more than 25 publications in reputed journals and conferences and has authored chapters in two books. She is currently working on BIG DATA, Information retrieval and Ontology. She is a recipient of Best paper Award in 2nd International conference on Recent Development in Sciences Engineering and Technology organised by GD Goenka University, Gurugram.