Application of Open-Source Software in Knowledge Graph Construction

Qianqian Cao\textsuperscript{1,2} and Bo Zhao\textsuperscript{1}\textsuperscript{\(\overset{\text{(E)}}{=}\)}

\textsuperscript{1} Yunnan Normal University, Kunming 650500, YN, China
569094105@qq.com, ykzb63@126.com
\textsuperscript{2} Key Laboratory of Educational Information for Nationalities, Ministry of Education, Kunming 650500, YN, China

Abstract. Knowledge graph (KG), as a new type of knowledge representation, has gained much attention in knowledge engineering. It is difficult for researchers to construct a high-quality KG. Open-source software (OSS) has been slightly used for the knowledge graph construction, which provide an easier way for researchers to development KG quickly. In this work, we discuss briefly the process of KGC and involved techniques at first. This review also summarizes several OSSs available on the web, and their main functions and features, etc. We hope this work can provide some useful reference for knowledge graph construction.

Keywords: Knowledge graph · Open-source software · Knowledge graph construction

1 Introduction

Knowledge graph (KG) is helpful to organize massive information efficiently. Although KG is applied in semantic search [1], intelligent question-answering [2] and decision-making widely, it’s too difficult for many people to construct knowledge graph and its method of automatic construction is not mature yet. However, there are a great many outstanding open-source softwares available and each has its own unique features. It is a key for developers how to choose an appropriate open-source software (OSS) to construct the knowledge graph. In the paper, we have analyzed and compared the characteristics of open-source tools which are applied in the process of knowledge graph construction (KGC), in order to give researchers a reference.

Firstly, we discuss briefly the general process of KG from the relevant literature. Secondly, some of the current OSSs for knowledge graph construction are analyzed in Sect. 3. Finally, the conclusion is given.

2 The Process of Knowledge Graph Construction

The term “knowledge graph” was first proposed to improve the quality of information retrieval by Google [3] in 2012. Since then, the term has been widely used without a
specific definition and representation of the knowledge based on graph could be con-
sidered a knowledge graph. In this paper, KG is a semantic graph consisting of vertices
(or nodes) and edges, while the vertices represent concepts or entities and the edges
represent relationships between the entities [4]. It able to organize information in an
easy-to-maintain, easy-to-understand and easy-to-use method. It has become prevalent
in education during these years. The process of knowledge graph construction includes
3 aspects: knowledge acquisition, knowledge fusion, knowledge processing, which will
be discussed as follows.

2.1 Knowledge Acquisition (KA)

KA is used to mine entities, entity attributes and relationships between the entities from
structured, semi-structured, even unstructured data sources. There are mainly three cor-
responding processing methods for three types of data respectively when knowledge
is acquired. For structured data, D2R [5] (an XML-based language) is used to map it
to RDF [6] schema [7] (a type of hierarchy which is used to define types and possible
relations). RDF (Resource Description Framework) is generally used to formalizes
structured information and present it graphically. Semi-structured data (e.g. encyclopedic
data, web data) can be gained automatically by using an unsupervised clustering
algorithm. For unstructured data, the information extraction technique is mainly used to
extract entities, entity attributes and relationships between the entities. Open information
extraction (OIE) [8] is the promising field in information extraction. OIE is based on
linguistic model and machine learning algorithm, which is used to extract information
from open domain. However, the OIE technique might lead to the lower precision and
lower recall rate. Through the process of KA, entities, attributes and relations between
entities are obtained from structured, semi-structured and unstructured data sources.

2.2 Knowledge Fusion (KF)

Through knowledge acquisition, it might contains lots of redundant and erroneous
information. KF focuses on re-cleaning and integrating those extracted results, such
as improving the connection density (the distance in the KG) of entities/relations and
making entity relations to present definite logic and levels. It mainly includes such task
as entity linking and knowledge merger. Entity linking focuses on the process of entity
disambiguation and entity resolution. Entity disambiguation is used to disambiguate a
polysemous entity mention or infer that two different mentions are the same entity, and
entity resolution is used to identify and link different manifestations of the same real
world entities in various records. Then, entities are mapped from an input text to corre-
sponding entities in a target knowledge base. So, entity linking is used to link the enti-
ties extracted from semi-structured and unstructured data sources to knowledge graph.
Knowledge merger is used to link the entities extracted from structured data sources (e.g.
external knowledge base data or relative database data) to knowledge graph. Through
knowledge fusion, the ambiguity of the concepts is eliminated and redundant and wrong
concepts are eliminated.
2.3 Knowledge Processing (KP)

The result provided by knowledge fusion is not yet a knowledge graph. Knowledge processing is used to transforming the result into the more structured and networked knowledge architecture. KP mainly includes two aspects such as schema construction and knowledge reasoning. Schema construction refers to design the classes/concepts, relations, functions, axioms, instances for the knowledge graph. There are three ways of constructing method: artificial, automated and semi-automated construction. Knowledge reasoning aims at generating new knowledge based on the existing knowledge in the knowledge graph through computer reasoning. There are three common ways to reason over knowledge graph: logic-based reasoning, graph-based reasoning, deep learning-based reasoning. Through knowledge reasoning, knowledge graph is further enriched and expanded.

3 On OSS of Knowledge Graph Construction

The term “open-source” is derived from the computer software industry. It can literally be interpreted as open-source code. In the paper, we will analyze and discuss the functions and characteristics of open-source software (OSS) in detail as follows. The commonly used OSSs in knowledge graph construction are shown as in Table 1.

Table 1. The commonly used OSSs in knowledge graph construction phases.

| OSS      | Knowledge acquisition | Knowledge fusion | Knowledge processing |
|----------|-----------------------|------------------|----------------------|
| Protégé  | /                     | /                | Yes                  |
| KAON     | /                     | /                | Yes                  |
| DeepDive | Yes                   | /                | Yes                  |
| GATE     | Yes                   | Yes              | /                    |
| KnowItAll| Yes                   | /                | /                    |
| Dedupe   | /                     | Yes              | /                    |
| Limes    | /                     | Yes              | /                    |
| SOFIE    | Yes                   | /                | Yes                  |

Protégé [9] is an open-source software developed by the Stanford University School of Bioinformatics Research Center, which is used for schema construction and knowledge reasoning (knowledge processing). It allows pluggable components, plug-ins, which can visualize knowledge and reason. It has friendly graphic interactive interface, uniform style, strong interactivity, adaptability and operability. For schema construction, it is the main functions of Protégé, which include the construction of concept classes, relationships, attributes, and instances of the schema. It is designed to hide the underlying details (the schema description language) from your users. It can also perform knowledge-based reasoning by clicking on the “Reasoner” button (on the menu bar) and provide users with reasoning result and explanation in detail.
KAON [10] is an open-source toolkit developed by the University of Karlsruhe and the Research Center for Information Technologies in Karlsruhe. It is often used to construct schema (knowledge processing). KAON implements the functions of schema construction by using KAON-API. KAON-API provides mechanisms for storing and editing schema, and also has enabled applications to access and process the schema. KAON’s constructing, storing, querying and retrieving schema operations are all performed based on graph. KAON’s graph-based operation is more intuitive and convenient than Protégé in the constructing process.

DeepDive [11] is an open-source system developed by Stanford University’s InfoLab laboratory, which is frequently used for information extraction (knowledge acquisition) and knowledge reasoning (knowledge processing). For knowledge acquisition, it uses “weakly supervised learning” algorithm [12] to extract structured relations from unstructured data, and then determines whether a specified relationship exists between entities. For knowledge processing, DeepDive is used to solve knowledge reasoning problems based on factor graphs, which is a type of probabilistic graphical model composed of variables and factors (definition of the relationships between variables in the graph). DeepDive uses probabilistic inference to estimate the probability whether the existing knowledge would be true, then determine whether it will be retained. As a complete knowledge extraction framework, DeepDive provides users with application code and an inference engine. Users only need to write application code and inference rules used during reasoning for their specific tasks [13].

KnowItAll [14] is another open-source software developed by the Turing Center at the University of Washington, which is used in knowledge acquisition widely. Training was conducted by unsupervised learning [15] and domain-independent operations. It has advanced performance than previous system in the tasks such as pattern learning and subclass extraction by using the search engine to perform extraction from massive Web data. Reverb and OLLIE are the first and second generation information extraction components of KnowItAll respectively. Reverb extracts the binary relationship directly from the English sentence without to pre-specify the relations. OLLIE system can extract information based on the “Syntax Dependency Tree”. It has better extraction results and the more optimized “Long-term Dependency” effect than Reverb based on text sequence. The distinctive feature of KnowItAll is its use of bootstrapping method that does not require any manually tagged training sentences [16].

GATE [17] (General Architecture for Text Engineering) is an open-source knowledge extraction system developed by the University of Sheffield, UK, which is capable of solving almost all problems encountered when processing text. Therefore, it can be used in information extraction (knowledge acquisition) and entity disambiguation (knowledge fusion). For knowledge acquisition, GATE offers users the functions such as data processing, rule definition and entity disambiguation through its natural language processing components, which lays the foundation for information extraction. Each component has the open interface, so that they can be called by other systems freely.

Limes [18] is an open-source entity linking framework developed by the DICE (Data Science) group at Paderborn University. More specifically, it is a link discovery framework for metric spaces (a set together with a metric on the set.), which is used in
knowledge fusion. It uses statistics, prefix suffixes, and location filtering to calculate the similarity rate between the entities, while the “entity pairs” mismatched will be filtered out. The specific characteristic of Limes is that it allows users to configure the rules of entity resolution flexibly.

Dedupe [19] is an open-source python library developed by Gregg Forest and Derek Ederalso, which is used for knowledge fusion. It uses machine learning to perform fuzzy matching, deduplication and entity resolution quickly on structured data. Specifically, users only need to label a small amount of data selected during the calculation process. All labeled data is clustered and grouped by a compound clustering method. Then the duplicate labeled data will be removed based on the calculation of similarity features [20] and machine learning models [21]. A salient feature of Dedupe is that it supports users using user-defined data types to label the data which are used for training models.

SOFIE is an open-source software about knowledge processing developed by the Max Planck Institute, which can also be used to extract information (knowledge acquisition). For entity fusion, it can be used to parse natural language documents, extract knowledge from them and link the knowledge into an existing schema, and perform disambiguation based on logical reasoning. For knowledge acquisition, it processes the document, splits the document into short strings. SOFIE’s main algorithm is completely source-independent. SOFIE’s performance could be further boosted by customizing its rules to specific types of input corpora. With appropriate rules, SOFIE could potentially accommodate other IE paradigms within its unified framework.

4 Conclusion

In this article, we have reviewed how the knowledge graph is constructed. We have discussed several the main OSSs available and the usage of these tools in the knowledge graph construction. As previously discussed, there are lots of open-source tools for constructing knowledge graph. We argue that different tool has its own features, such as Dedupe is more suitable to process structured data, DeepDive has better performance when it is used to extract sophisticated relationships between entities, SOFIE integrates information extraction with logical reasoning firstly. To sum up, KGs have arisen as one of the most important tool for knowledge representation, storage and organization. It is hope that more open-source softwares will be developed to provide researchers more convenient ways to build, update, retrieve, and maintain knowledge graph.

Acknowledgments. The research is supported by a National Nature Science Fund Project (No. 61967015), and specific project of teacher education of Yunnan Province education science planning (Union of higher education teachers) (GJZ171802).

References

1. Grainger, T., AlJadda, K., Korayem, M., et al.: The Semantic Knowledge Graph: A compact, auto-generated model for real-time traversal and ranking of any relationship within a domain. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 420–429. IEEE (2016)
2. Zhang, Y., Dai, H., Kozareva, Z., et al.: Variational reasoning for question answering with knowledge graph. In: Thirty-Second AAAI Conference on Artificial Intelligence, AAAI (2018)

3. Amit, S.: Introducing the Knowledge Graph: things, not strings http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html. Accessed 1 March 2020

4. Yan, J., Wang, C., Cheng, W., Gao, M., Zhou, A.: A retrospective of knowledge graphs. Front. Comput. Sci. 12(1), 55–74 (2018). https://doi.org/10.1007/s11704-016-5228-9

5. Bizer, C.: D2r map a database to rdf mapping language (2003)

6. RDF Working Group RDF. http://www.w3.org/RDF. Accessed 1 March 2020

7. Paulheim, H.: Knowledge graph refinement: a survey of approaches and evaluation methods. Semant. Web 8(3), 489–508 (2017)

8. Banko, M., Cafarella, M. J., Soderland, S., et al.: Open information extraction for the web. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp. 2670–2676. New York (2007)

9. Musen, M.A.: The protégé project: a look back and a look forward. AI matters 1(4), 4–12 (2015)

10. Bozsak, E., Ehrig, M., et al.: KAON — Towards a large scale semantic web. In: Bauknecht, K., Tjoa, A., Quirchmayr, G. (eds.) EC-Web 2002. LNCS, vol. 2455, pp. 304–313. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45705-4_32

11. Zhang, C.: Deepdive: A Data Management System For Automatic Knowledge Base Construction. University of Wisconsin-Madison, Madison, Wisconsin (2015)

12. Zhang, Z.: Weakly-supervised relation classification for information extraction. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, pp. 581–588 (2004)

13. Mallory, E.K., Zhang, C., Ré, C., Altman, R.B., et al.: Large-scale extraction of gene interactions from full-text literature using DeepDive. Bioinformatics 32(1), 106–113 (2016)

14. Etzioni, O., Cafarella, M., Downey, D., et al.: Unsupervised named-entity extraction from the Web: an experimental study. Artif. Intell. 165(1), 91–134 (2005)

15. Zigoris, P., Eads, D., Zhang, Y.: Unsupervised learning of tree alignment models for information extraction. In: Sixth IEEE International Conference on Data Mining-Workshops (ICDMW’06), IEEE, pp. 45–49 (2006)

16. Etzioni, O., Cafarella, M., Downey, D., et al.: Web-scale information extraction in knowitall: (preliminary results). In: Proceedings of the 13th International Conference on World Wide Web, pp. 100–110 (2004)

17. The University of Sheffield. GATE Information Extraction. http://gate.ac.uk/ie/. Accessed 2 March 2020

18. Data Science Group at UPB. Limes. https://github.com/dice-group/LIMES. Accessed 2 March 2020

19. Gregg, F., Derek, E. Dedupe, https://github.com/datamade/dedupe, Accessed 3 March 2020

20. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Anal. Mach. Intell. 24(3), 301–312 (2002)

21. Wikipedia contributors. Machine learning. https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=944212235. Accessed 6 March 2020

22. Suchanek, F., Sozio, M., Weikum, G.: Sofie: a self-organizing framework for information extraction. In: Proceedings of the 18th International Conference on World Wide Web, pp. 631–640 (2009)