Telecom Business Review
13 (1) 2020, 20-26
http://publishingindia.com/tbr/

A Survey of Methods, Tools and Applications of Knowledge Base Construction (KBC)

Devesh Rajadhyax

Founder and CEO, Cere Labs Pvt. Ltd., Maharashtra, India. Email: devesh.rajadhyax@cerelabs.com

Abstract

Knowledge Bases (KBs) have recently become very valuable because of their use in many Artificial Intelligence (AI) applications. For example - KBs are a critical part of conversational agents that are rapidly being adopted by the industry. Knowledge Bases are made up of a number of facts about the real world, and the number of such facts is typically very large. The construction of the KB involves identifying the facts from unstructured data such as text, images, videos and speech. Due to the challenges of processing unstructured data, Knowledge Base Construction (KBC) used to be done manually, requiring huge efforts and long timelines. In recent years, intelligent technologies such as Machine Learning and Deep Learning are being employed for the purpose of KBC. In this paper, we provide an introduction to KBC and describe how various cutting edge technologies are being employed for its automation. We then mention some KBC systems created by academic as well as commercial groups. We survey some solutions in the industry that are already using KBC. In addition, we attempt to predict the future research possibilities in this field.

Keywords: Knowledge Base Construction, Information Extraction, Text Processing, Knowledge Graph, Machine Learning

1. Introduction

The last few years have seen unprecedented growth in intelligent systems such as Artificial Intelligence (AI). AI systems aim at developing human capabilities through learning from data. A 'learner' that performs this task is the central component of any AI system. Learners are implemented using various algorithms, but a large set of real world knowledge is always required to prepare this learner. A logic based learner such as decision tree is almost entirely dependent on a set of facts about the world. A learner using Artificial Neural Network (ANN) requires huge amount of knowledge to train its network. Probabilistic Graphical Models such as Bayesian and Markov use knowledge as priors as well as for grounding the network.

There are many reasons why AI systems use knowledge other than for learning. A conversational agent (also called chatbot) requires knowledge for answering questions. Augmented Analytics systems use the knowledge to map user needs on system queries. This growing list of applications has made Knowledge Bases (KBs) a vital part of AI architectures (Ratner & Re, 2018).

Knowledge Bases store huge amount of facts about the world. The architecture of KBs make it easy to store various types of facts and query them. The knowledge of the world resides in a variety of sources, most of them unstructured. There are text sources, such as documents, websites and emails and non-text sources such as images and videos. The text itself may be plain or formatted in tables and charts. Knowledge has to be extracted from all these sources and populated in the KB to be useful.

Considering the above points, it is then easy to imagine that the process to populate a KB is of utmost importance. This process is called Knowledge Base Construction (KBC). A KBC system must be able to understand and extract relevant facts from a large variety of sources. The complexity and uncertainty in the source material make KBC a challenging task. The performance of a KBC system is measured by the accuracy (how much knowledge was correct) and completeness (what fraction of knowledge contained in a source was extracted).

This paper is an introduction to KBC. The author has been working on various aspects of KBC during the development of a Learn and Recite (L&R) software system at his organization. The material given here is the result of studies done for the conceptualization and design of this system over past few years. The objective of this paper is to introduce students and engineers to the emerging and promising field of KBC and to facilitate research and development in the area.

The paper is organized as follows – Section 3 explores the history of Knowledge Base Construction, Section 4

enlists the challenges faced by KBC systems, Section 5 delves into the architecture of the systems in terms of components and technologies used, Section 6 looks at some of the existing KBC systems, Section 7 mentions the applications of KBC in the industry.

2. Literature Review

In Knowledge base construction is the process of populating a knowledge base (KB) with facts extracted from various types, especially unstructured data (Computer Science Department, Stanford University [CS-SU], n.d,), (Shin et al., 2016). The challenges in this process due to the unstructured nature of data and the various subtasks involved have been well documented in work such as Ratner and Re (2018). The recent interest in knowledge base construction due to QA agents, also called chatbots has been described by Subasic, Yin and Lin (2019).

While the author derives the various steps involved in KBC from his own experience in the field, the principals involved in the tasks are referred from the excellent sources provided by leading researchers in the field, such as Oro and Ruffolo (2009), Eftimov, Seljak and Korosec (2017), Zhang and Wang (2015). The applications and existing systems have been studied from the original papers published by the creators of the systems, like Dong et al. (2014).

3. History

The origins of knowledge bases can be traced to the 'Expert Systems' of 1970s that were probably the first commercially successful application of AI (Leonard-Barton & Svikova, 1998). The expert systems phase convinced some scientists about the importance of collecting real world knowledge. The Cyc project started in 1984 with an objective of collecting Common Sense Knowledge – knowledge about real world entities and the rules binding them together (Lenat, 1995). Cyc is the world's longest running knowledge collection program and contained more than 1.5 million terms as of 2017.

In the first decade of 2000's, the concept of semantic web drew attention of the industry and scientific community. Semantic web technologies such as RDF and OWL have since become important parts of the KB technology (Gangemi, 2013). The 2011 win of Jeopardy! competition by IBM Watson became a good demonstration of the power of knowledge bases and their use for question answering.

In the 2010's, two factors became the main drivers of knowledge base technology. Since the 2010 introduction of Siri in Apple Inc's iOS, virtual assistants have gained prominence. The assistants and their variation the chatbots have found commercial success (Ratner & Re, 2018). Many voice enabled assistants such as Google Assistant, Amazon's Alexa, Microsoft's Cortana are being widely used by people. Since these assistants are primarily question answering agents, knowledge base is an integral part of their architecture.

The second driving factor is the rise of Deep Learning as the preferred method of implementing AI. While DL is a powerful learning method, it requires huge amount of labeled data for preparing the learner. Automated knowledge base construction from variety of sources has emerged as the standard solution to this requirement.

4. Challenges

Constructing a knowledge base is a challenging task. The major challenges arise because of three reasons:

4.1. Nature of Sources

By its very definition, KBC involves processing of unstructured sources to generate structured knowledge. Unstructured sources include text, scanned documents, pdf's, web pages, images and so on (Computational Biology Institute [CBI], 2018). These sources are difficult to read for computer programs due to following of their characteristics (Ratner & Re, 2018), (Lin, Liu, Sun, Liu & Zhu, 2015):

Uncertainty: Various layouts, use of elements like tables and charts and rich formatting introduce element of doubt in the interpretation of most unstructured sources. Even when the source is pure text, the meaning of the text is ambiguous due to the nature of natural language. Scans and images contain noise that adds further uncertainty.

*Complexity***:** Knowledge is a combination of various elements such as entities, relationships, intents, purposes and so on. The number of combinations of these elements is very large, posing a challenge for the systems that seeks to identify only those combinations that are both meaningful and useful.

Configuration Efforts: The KBC system is owned by domain experts. If the system is too complex and involves technical activities, the domain experts have to work with technical experts for the configuration. This makes the configuration too expensive and difficult to organize. On the other side, the system should be powerful enough to require minimum manual efforts to create the KB. Balancing the simplicity in configuration with the power of automating knowledge extraction creates the effort challenge.

Quality Requirements: The KB is used for a specific purpose such as training a machine learning model or answering questions. In order to fulfil this purpose, the generated KB must satisfy certain quality criteria (Subasic et al., 2019). The main requirements are:

Completeness: Out of the total knowledge required for the job, how much has been harvested (Razniewski, Suchanek & Nutt, 2016).

Accuracy: Out of the total knowledge harvested, how much is accurate (Subasic et al., 2019), (Razniewski, Suchanek & Nutt, 2016).

For example, for a chatbot, the knowledge should be such that it can provide correct answers to all expected answers. These quality requirements put a challenging responsibility of the KBC system.

5. Architecture of KBC Systems

The knowledge contained in a KB can be classified as schema elements (entities and relationships) and data elements (instances). While data elements represent the ultimate use of the KB, schema elements are used for reasoning and selection of the right data elements for the purpose. For example, in a pharma KB, the categories and properties of molecules are part of schema, whereas the trial results are instances.

The schema elements are usually found in natural language format ('Tobramycin is an aminoglycoside antibiotic'). Data elements are found both in natural language and in formatted layout such as a table. For example, the test results are more likely to be found in a table than in a paragraph of text.

The extraction from text involves Natural Language Processing (NLP). NLP based knowledge extractors can use grammatical analysis implemented through regular expressions and FSA based algorithms. The grammatical method is fast and effective, however it involves complex programming and has limited scalability. In the last few years, Multi-layer Neural Networks (MLNN), termed as Deep Learning (DL) models are increasingly being used for text processing. These models require sizeable amount of data to get off the ground, but they 'democratize' the KBC process as non-programmers can also train and use DL based extraction pipeline.

The data elements are mainly found in formatted and image-based sources such as scanned documents and pdf's. The pipeline to extract instances from these sources has to utilize spatial relationships and contains modules to convert image to text (popularly termed as OCR), analyze the layouts and tag the values with appropriate schema element (Oro & Ruffolo, 2009), (Oro & Ruffolo, 2008).

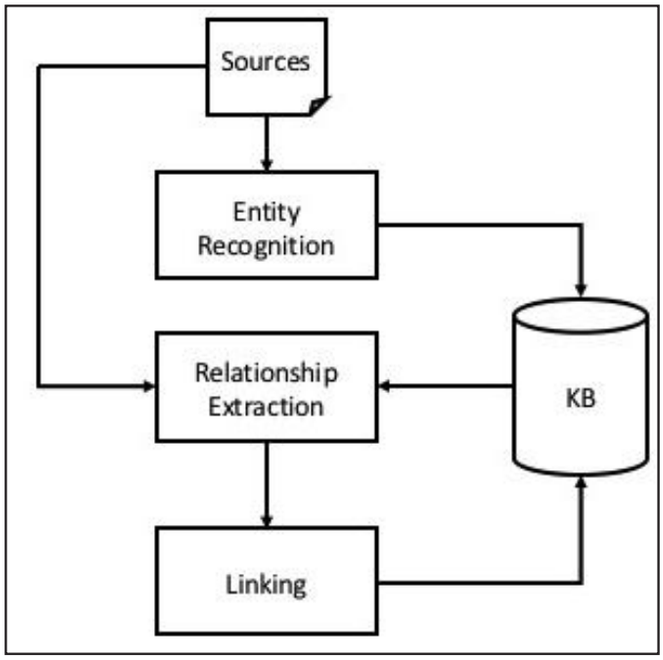

Fig 1: Basic Pipeline of Knowledge Extraction

Some of the important tasks in both pipelines are explained here:

Entity Identification: Entities are usually noun phrases. They can be identified by various methods in Computational Linguistics (CL) such as Part of Speech (POS) tagging and Named Entity Recognition (NER) (Eftimov et al., 2017). Compound entity identification is a recursive process that involves identifying both simple entities and relations.

Relationship Extraction: Extracting relationships is a tough challenge owing to the large number of possible relationships and also the large number of possible ways in which they can be expressed in text. To counter this, some methods (Clark et al., 2014) require the relationships to be defined in advance to narrow the possibilities.

Various methods have been employed to identify and extract relationships. Use of regular expressions to identify patterns is popular since early days of KBC (Clark et al., 2014). More advanced NLP tools such as Dependency Parsers are used by some systems (Al-Zaidy, Rabah & Giles, 2018). In recent times, Deep Learning (DL) models are used to find relation mentions (Zhang & Wang, 2015), (Santos, Xiang & Zhou, 2015) and to extract the relation.

OCR: OCR extracts text from image sources. Older OCR tools used image processing and had limited accuracy. Modern OCR engines use DL models and are proving to be far superior (Breuel, 2008), (Breuel, 2017).

Layout Analysis: As mentioned, instance data is usually found in tables or charts. Even when the table is not explicitly drawn, the elements of data are arranged on the paper as though they belong to a table. Extracting such data requires spatial understanding of the document to correctly group the text returned by OCR (Oro & Ruffolo, 2009). This is an area of continuing research and various methods that use different ML/DL models are being tried out.

Schema Tagging: Extracting data elements involves associating them with the right schema elements. For instance, if the viscosity value of a drug is mentioned in a specification table, it has to be tagged as the value for the property viscosity for the particular drug, along with probably its batch number. Since many layouts are ambiguous, this is a challenging task requiring resolution and judgment. A host of techniques are being developed that employ some form of uncertain inference.

6. Existing Systems

The Knowledge bases and KBC systems currently available broadly fall into three categories:

6.1. Publicly Available KBs

The best known example of these systems is DBPedia (Lehmann et al., 2012), which extracts structured information from Wikipedia pages and makes it available as RDF dataset over the World Wide Web. The DBPedia extraction process uses the 'infoboxes' included in Wikipedia pages and thus is more focused on relationship extraction and linking process. The current release of DBPedia has 4.5 million entities in its English dataset. Another example of a public KB is the Never Ending Language Learning (NELL) project (Mitchell et al., 2018). NELL is a KBC system that extracts facts from the web. An initiative of the Carnegie Melon University, it is running continuously since January 2010 and has collected around 50 million candidate facts in its knowledge base thus far.

6.2. Open Source Platforms

Quite a few KBC platforms were developed in academic institutions and were subsequently made available to the community as open source software. Deep Dive is a KBC platform developed in the Stanford University (Shin et al., 2015). Its commercial version was acquired by Apple Inc in 2017. Snorkel is another system created by Stanford that focuses on creating weakly supervised data for training deep learning models (Ratner et al., 2017). Fonduer, also developed at Stanford, aims at extracting knowledge from richly formatted text (Wu et al., 2018).

6.3. Commercial Products

Many technology companies have developed software products that have KBC capabilities. The leader in KBC systems is IBM, whose Watson set the KBC trend rolling in 2011. IBM has developed another product called Socrates that powers semantic search through knowledge base extracted from text (IBM Research Editorial Staff [IBM-RES], 2017). SystemT is another product by IBM that is used for quite a few commercial KBC projects (Li, Reiss & Chiticariu, 2011). Google has its Knowledge Vault (Dong et al., 2014) that is used to power its 'information boxes' in search and to answer queries to Google Assistant.

7. Applications in Industry

The Knowledge bases and knowledge base construction systems are being employed by many commercial organizations across the world. While some are using commercially available products, others are using open source tools to build their own KBC systems. Here we try to explore the usage of KBC in some noteworthy applications:

7.1. Search

As the volume of data has multiplied in recent years, search has become an important application both for individuals and organizations. Internet search which is probably the most used application of all time has evolved from being a directory lookup to handling queries semantically. Google search recognizes entities being searched and shows detailed information using its Knowledge Vault (Dong et al., 2014). Wolfram Alpha is a search engine designed to handle semantic queries (Johnson, 2009).

Enterprise search is an equally important area considering that huge organizations own truly gigantic quantities of data. Products such as IBM's Watson Discovery help to create search indexes that can be used for querying. Socrates, a KBC product from IBM product can be used along with Watson Discovery to populate a knowledge base (IBM-RES, 2017).

7.2. Virtual Assistants

Siri, introduced by Apple Inc. in its iPhone in 2011 was the first well known virtual assistant. Since then Google Assistant by Google, Cortana by Microsoft, Alexa by Amazon and other assistants have been introduced. Special devices such as Amazon Echo and Google Home have become popular. They are all supported by large knowledge bases of their own (Ratner & Re, 2018), apart from using public KBs like Wikidata (Simonite, 2019). The construction method of most of these organizations is not known.

Chatbots are virtual assistants that can handle natural language text as input. Most of the chatbot providers also include a KB and some KBC mechanism to populate it. Rasa, a widely used open source chatbot platform, can be integrated with a knowledge base such as Grakn (Bocklisch, Faulker, Pawlowski & Nichol, 2017), (Bergmann, 2019).

7.3. Research Databases

KBC is being used to extract data from a large number of sources for research purposes. Deep Dive and Snorkel, the KBC systems developed in the Stanford University have been used to create databases for paleontology record, human trafficking and pharmacology (Computer Science Department, Stanford University [CS-SU], n.d).

7.4. Training Data Generation

Manually labelling data for training DL models is expensive and time consuming. Weak supervision is a technique of training a DL learner that uses small quantities of good quality labelled data combined with large amounts of weakly labelled data (Ratner et al., 2017). Weakly labelled data contains labels that are not 100% accurate. KBC is used to generate such weakly labelled data, reducing the time and cost of the model training activity.

8. Conclusion and Future Directions

In this paper we introduced the knowledge base extraction (KBC) systems that have become very important constituents of AI solutions. KBC systems face challenges due to unstructured nature of sources and quality requirements on the knowledge extracted. There is a growing need to make KBC accessible to domain experts rather than programmers. This need to 'democratize' KBC is driving larger use of deep learning based architectures. Since an estimated 80% of the world's data is in unstructured form, the demands on KBC systems are going to increase in coming times. Future research in this field must focus on building learning based system that can be trained by domain experts on an ongoing basis. Since KBC can also create training data for machine learners, such architectures can give rise to an endless virtuous loop of knowledge creation.

9. Implications of Research

Industry persons and researchers working on applications such as chatbots, automation, text processing, analytics are already employing knowledge bases and knowledge graphs to a large extent. They will be able to use the material presented in this paper to take advantage of KBC for populating their knowledge bases. A fundamental understanding of the working principles of KBC will enable to make better utilization of the existing KBC systems such as mentioned in this paper more effectively.

References

Al-Zaidy, R. A., & Giles, C. L. (2018). Extracting semantic relations for scholarly knowledge base construction. *2018 IEEE 12th International Conference on Semantic Computing (ICSC)* (pp. 56-63).

Bergmann, T. (2019). Setting up a knowledge base to encode domain knowledge for Rasa. *Medium*. Retrieved from https://blog.grakn.ai/setting-up-a-knowledge-base-t o-encode-domain-knowledge-for-rasa-6c936242d03d

Bocklisch, T., Faulkner, J., Pawlowski, N., & Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. *ArXiv, abs/1712.05181.*

Breuel, T. M. (2008). The OCRopus open source OCR system. *Electronic Imaging*.

Breuel, T. M. (2017). *High performance text recognition using a hybrid convolutional-LSTM Implementation.* 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (pp. 11-16).

Clark, P., Balasubramanian, N., Bhakthavatsalam, S., Humphreys, K., Kinkead, J., Sabharwal, A., & Tafjord, O. (2014). *Automatic construction of inference-supporting knowledge bases.* 4th Workshop on Automated Knowledge Base Construction (AKBC).

Computational Biology Institute. (2018). Workshop: Rapid biomedical knowledge base construction from text. Retrieved from https://cbi.gwu.edu/workshop-rapid-biomedical-knowledge-base-construction-text

Computer Science Department, Stanford University. (n.d). DeepDive applications. Retrieved from http://deepdive.stanford.edu/showcase/apps

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., & Zhang, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. *KDD'14*.

Eftimov, T., Seljak, B. K., & Korosec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS ONE, 12*.

Gangemi A. (2013) A comparison of knowledge extraction tools for the semantic web. In Cimiano P., Corcho O., Presutti V., Hollink L., Rudolph S. (eds) *The Semantic Web: Semantics and Big Data* (pp. 351-366). ESWC 2013. Lecture Notes in Computer Science, vol. 7882. Springer, Berlin, Heidelberg.

IBM Research Editorial Staff. (2017). Automated knowledge base construction solution wins at ISWC 2017. Retrieved from https://www.ibm.com/blogs/research/2017/11/knowledge-base-construction-iswc-2017/

Johnson, B. (2009). British search engine could rival Google. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2009/mar/09/search-engine-google

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Kleef, P. V., Auer, S., & Bizer, C. (2015). DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web, 6*, 167-195.

Lenat, D. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM, 38*(11), 33-38.

Leonard-Barton, D., & Svikova, J. (1988). Putting expert systems to work. *HBR*. Retrieved from https://hbr.org/1988/03/putting-expert-systems-to-work

Li, Y., Reiss, F., & Chiticariu, L. (2011). System T: A declarative information extraction system. *ACL*.

Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. *AAAI*.

Mitchell, T., Kisiel, B., Krishnamurthy, J., Lao, N., Rivard, K., Mohamed, T., Nakashole, N., Platanios, E. A., Ritter, A., Samadi, M., Settles, B., Cohen, W., Wang, R., Wijaya, D., Gupta, A., Chen, X., Saparov, A., Greaves, M., Welling, J., & Gardner, M. (2018). Never ending learning. *Communications of the ACM, 61*(5), 103-115.

Oro, E., & Ruffolo, M. (2008). Towards a system for ontology-based information extraction from PDF documents. R. Meersman & Z. Tari (eds), On the move to meaningful internet systems: OTM 2008. *Lecture Notes in Computer Science,* vol. 5332. Springer, Berlin, Heidelberg.

Oro, E., & Ruffolo, M. (2009). *PDF-TREX: An approach for recognizing and extracting tables from PDF documents.* 10th International Conference on Document Analysis and Recognition (pp. 906-910).

Ratner, A., Bach, S. H., Ehrenberg, H. R., Fries, J. A., Wu, S., & Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases, 11*(3), 269-282.

Ratner, A., Ré, C., & Bailis, P. (2018). Knowledge base construction in the machine-learning era. *Communications of the ACM, 61*(11), 95-96.

Razniewski, S., Suchanek, F. M., & Nutt. W. (2016). But what do we actually know? *AKBC*.

Santos, C. N., Xiang, B., & Zhou, B. (2015). Classifying relations by ranking with convolutional neural networks. *ArXiv, abs/1504.06580*.

Shin, J., Wu, S., Wang, F., Sa, C.D., Ratner, A., Zhang, C., & Ré, C. (2016). Incremental knowledge base construction using DeepDive. *The VLDB Journal, 26*, 81-105.

Subasic, P., Yin, H., & Lin, X. (2019). *Building knowledge base through deep learning relation extraction and Wikidata.* AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering.

Tom Simonite. (2019). Inside the alexa-friendly world of wikidata. Retrieved March, 2019, from https://www.wired.com/story/inside-the-alexa-friendly-world-of-wikidata/

Wu, S., Hsiao, L., Cheng, X., Hancock, B., Rekatsinas, T., Levis, P., & Ré, C. (2018). Fonduer: Knowledge base construction from richly formatted data. *Proceedings of the 2018 International Conference on Management of Data*.

Zhang, D., & Wang, D. (2015). Relation classification via recurrent neural network. *ArXiv, abs/1508.01006*.