Survey on Generic and Biomedical Knowledge Graph

Ekant Amin, Pushpak Bhattacharyya
Indian Institute of Technology, Bombay
{ekant, pb} @cse.iitb.ac.in

Abstract

In this era where we have an immense amount of information that we are generating, every application, be it large or small, relies on the availability of a massive amount of data to dig upon, find statistics through it and present them to user with a certain confidence level. However, all of this new information comes with its own implications. One of the major concern with having such huge data warehouses is the unavailability of the ability of storing the data in an explainable manner. Knowledge Graph provides a semantically aware mechanism of doing so and has proved its worth in a number of domains like medical, music, politics etc. However, the idea behind this is not new and became popular back in 1980s with the introduction of "WordNet". Despite being a relatively older concept and a number of people working on it, creating a knowledge graph still is a tedious task. In this paper, we provide basics of knowledge graph and its components. We have also introduced existing knowledge graphs - both generic and domain-specific, biomedical being the domain of our concern. Finally, we have included some of the works done in this domain that highlight the creation techniques for a knowledge graph.

1 Introduction

Tasks involving Natural Language Understanding inevitably require the power of semantically understanding the information and not only capturing the syntactic aspect of it. For semantic understanding of the information, the information needs to be carefully and logically organized in a structure that clearly represents the different entities involved in the task and the relations between them. Knowledge Graph is one such structure. This survey aims at exploring literature available for the work that has already been done in the domain of developing a knowledge graph pertaining to medical and health area. We have tried to summarize existing knowledge bases - both generic and specific to medical domain & various techniques of ontology learning.

2 Knowledge Graph

A **Knowledge Graph** is a graph where each node or each vertex represents an entity from the real world and each edge represents the relation between two entities that are connected by that particular edge.

Symbolically, these entities are represented using oval designs while the relations are represented using directed arrows. These entities and relations are then populated with instances from the data at hand to build a compact representation of the otherwise raw-form data. This concept of entities and instances is very similar to the concept of Classes and Objects in Object Oriented Programming Paradigm. Usually, these knowledge graphs are populated with huge amounts of data and upon doing so, they act as hub of information encoding semantic knowledge in clear and concise manner.

A simple example of entity and relation is " President is a "Politician". Here "President" and "Politician" are the two entities which are connected by "Is A" relationship. Notice that the relation between the two entities is a directed one. That is, every relation has a "Domain" and "Range". In this particular example, entity "President" being the domain and entity "Politician" being the range of the relation "Is A".

An entity can be connected to multiple different entities via such directed edges effectively resulting in a directed multigraph and since each relation is depicted as a directed edge, this representation provides a more intuitive way of both querying the data and understanding the returned results.

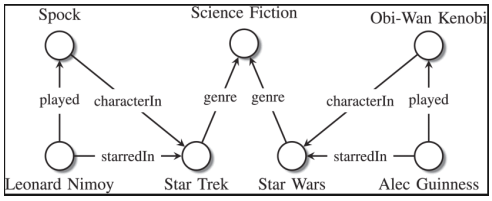

Figure 1: Illustration of Knowledge Graph (Nickel et al., 2016)

2.1 Concept of Triples

A common way of representing relationships in a Knowledge Graph is in the form of Triples (Subject, Predicate, Object) where Subject comes from the domain, Predicate comes from the relation set and Object comes from the range of the relation. One instance of such triple can be (President, Is A, Politician). Such heterogeneous relations can be incorporated in the Knowledge Graph and enable it to link far-away entities in a meaningful and clear way.

Knowledge Graphs contain tremendous amount of information in them, however like any other knowledge base, it can never be completely trusted for completeness. According to (Nickel et al., 2016) Non-existing Triples can be interpreted in two different ways:

Closed World Assumption(CWA)

Under this assumption, the non-existing triples are treated as "False", effectively meaning that the entities that are not related in Knowledge Graph are not related at all.

Open World Assumption(OWA)

Under this assumption, the non-existing triples are treated as "Unknown", that is in case two entities are not related in a Knowledge Graph, it does not mean that they are not related. In fact, they can be related and just that we don't have any evidence of their relationship given the available data.

2.2 The Complete Picture

A simple Knowledge Graph depicting all the properties stated above can be shown below:

A typical knowledge graph would have following characteristics:

- **Declarative**: A Knowledge Graph always has a meaning encoded with its components and is independent of the platform and the algorithm it is implemented with.
- **Non-hierarchical:** It is not a tree. It can be, however the typical amount of complexity of data for which a knowledge graph is created cannot be simply represented in a tree structure.
- **Annotated:** A knowledge Graph does not only represent the entities and relations, but can also contain annotations and meta data for its components thus making it an enriched source of context.
- **Large:** A typical knowledge graph can contain millions of nodes and can still be incomplete.

2.3 Applications

Knowledge Graphs are now inherent part of all popular applications in today's world, ranging from recommendation systems to conversational systems, even including Alexa, Siri, Google Assistant and other popular voice assistants.

- These Knowledge Graphs can map the questions that we ask to our voice assistant to an organized set of information and help in acheiving better quality answers to it. In a way, with the help of human-encoded information, these graphs can add common sense to the systems they augment.
- When used effectively, Knowledge Graph can be used to represent the

meaning and relationship between entities. This representation is usually termed as Ontology. These ontologies are periodically updated as and when new data arrives.

- Organizations, these days, have tremendous amount of data available at their disposal and this data comes from multiple sources. A human being can seldom make use of his abilities to make sense out of such huge amount of data. However, Knowledge Graph can be used to organize such data in a more user-friendly and readable form. Hence, in a way, effective content management is a huge advantage of using a Knowledge Graph.

- In today's world, when Deep Learning is at its uprecedented growth rate, there is one thing that bothers almost every ML researcher and that is the black box nature of these deep learning systems. These systems are not able to explain how and why they resulted in a particular ouput, hence there is a severe need for Explainability. This need is catered by Knowledge Graphs. If the decisions flows from a deep learning system can be documented using a knowledge graph, the decision making process could be made more transparent.

- Another major use case of a knowledge graph is in text-based search engines. Search results can be enhanced with semantic information from the knowledge graphs. Such transformations can help lead us from normal text-based search engines to a question-answer service with semantic awareness.

2.4 Challenges with Knowledge Graph

One of the most challenging aspect of a Knowledge Graph is its size. A knowledge graph with millions of vertices is not something that is unrealistic and does not exist. Such a huge graph has considerably high memory requirements to function and could even require hundreds of gigabytes of working memory.

An even bigger challenge is the computational cost involved while querying such a huge graph. Even the most basic graph queries like the length of path between two particular nodes can prove to be computationally very expensive for a huge graph.

Cache coherence is one possible solution and currently-in-use approach to lower down the computation time.

3 Existing Knowledge Graphs

3.1 Wordnet

Wordnet (Miller, 1995) is the most popular dataset which stores semantic information of lexical words. It is the knowledge base which can be considered as a thesaurus as it provides synonyms and meaningful information about the word.

Wordnet stores a word along with its sense. Meaning of the word is defined by its sense. So the words having same sense are synonymous. Words having more than one sense are polysemous. Wordnet also stores the context in which the word can appear. This set of contexts is named as Set C. C is further partitioned based on syntactic categories like Noun, Verb etc. Subset N stores nouns. Subset V stores verbs.

Relationship between the words are defined between the senses. Examples of Semantic relations are shown in figure 2.

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry

Note: N = Nouns Aj = Adjectives V = Verbs Av = Adverbs

Figure 2: Semantic relations of Wordnet (Miller, 1995)

- Synonymy: It is the relation that joins synonyms. A synset in wordnet contains words with same sense. It is a symmetric relationship between two sense.

- Antonymy: This is also a symmetric relation. For example: good and bad are the antonyms of each other. Antonyms are generally adverbs and adjectives, so it is useful in describing them.

- Hyponymy and Hypernymy: These relations are transitive in nature and inverses of each other. They are subnames and supernames respectively describing specific and generic sense. For example: dog is an animal, dog is the hyponymy and animal is the hypernymy. It is useful in describing nouns.

- Meronymy and Holonymy: These two relations are also inverses of each other. They describe the "part-of" relationship. For example: motherboard is a part of CPU, here 'motherboard' is Meronymy and 'CPU' is Holonomy.

- Trophonymy: It is same as meronymy but used for describing verbs. As number of verbs are comparatively lesser than nouns, hierarchy for the same is also shallower.

- Entailment: This relation is described between verbs. Entailment is "if A then there must be B". An example for this: He is driving , then it is also sure that he is riding. Here the verb 'drive' entails the verb 'ride'.

These relationships are the pointers between senses or word forms.

3.2 Freebase

Freebase (Bollacker et al., 2008) is the database where World's knowledge(general human knowledge) is stored in structured way in collaborative manner. This database is an open-source. Many Communities are involved in the creation of Freebase.

Inspiration behind the freebase model is publicly available knowledge sources like Wiki and Semantic Web. Wiki is semi-structured knowledge created by collaborative approach, which makes it diverse and involves heterogeneous data, but supports very less structured query aided tools.

Freebase has tried to capture both, structures database's scalability as well as wiki's collaborative approach. As mentioned in /ref-paper, in freebase data, number of types are more than 4000, number of tuples are more than 125,000,000 and number of properties are more than 7000.

Some important features of Freebase are as follows:

- Storage of tuple is scalable in Freebase. It provides undo feature as a built-in mechanism and lets people collaboratively add data on large scale and maintain them. This is how Freebase has got the huge dataset it currently has.

- For reading and writing purpose, an Application Programming Interface(API) was provided, which was supporting MQL(MetaWeb Query Language). However, this API was taken down in 2015.

- To make it lightweight and non-rigid, Freebase provides a Type-system which has no natural convention. Different users can make entities of different types depending on their beliefs and interpretation. This, however, has led to some conflicting types in existence for the same entity. However that has been done to demonstrate differenceof opinions the users have.

- For identifying one entity from real world, freebase assigned only one GUID resulting in complete normalization.

3.3 ConceptNet

It (Speer et al., 2017) is a semantic network for words and is freely available and developed under Open Mind Common Sense

project, MIT. It is crowd-sourced project. Assertions in the conceptnet are in the form of triples. for example "A car has an engine". From this sentence, a triple can be extracted that signifies the relation between Car and Engine : (car,HasA,Engine).

Additionally, ConceptNet also provides links to the definitions from other datasets also i.e. for a particular word, it will have its own definition and also have links to definitions from other sources like DBpedia and wiki . ConceptNet is useful in NLP learning techniques. Word-embeddings based on contextual similarity can certainly be aided with the information of word relatedness that is provided by ConceptNet.

4 Bio-Medical Knowledge Resources

4.1 Unified Medical Language System(UMLS)

There are many biomedical vocabularies available for use but they are for specific purposes. UMLS (Bodenreider, 2004) has combined them all to create a single resource where all the concepts or entities can be identified with their unique code. It has also mapped entities between different resources. These resources with their specific entities are mentioned below and shown in figure 3.

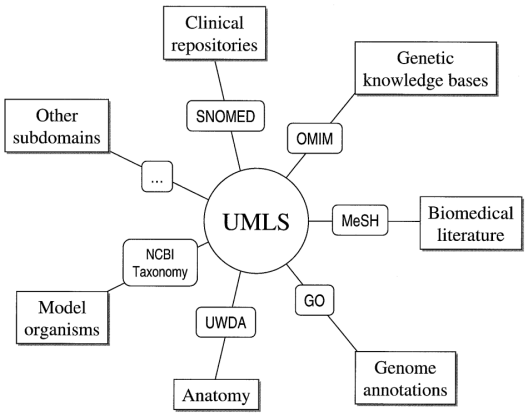

Figure 3: The various sub-domains integrated in the UMLS (Bodenreider, 2004)

- **SNOMED CT** for clinical repository
- **OMIM** for genetic knowledge bases
- **MeSH** for biomedical literature
- **GO** for genome annotations
- **UWDA** for anatomy
- **NCBI Taxonomy** for model organisms

UMLS consists of three major parts:

- Metathesaurus : Vocabulary of all the collected terms of medical domain.
- Semantic Network : Categorization of terms and relationships between them.
- SPECIALIST Lexicon and Lexical tools : Normalization of terms.

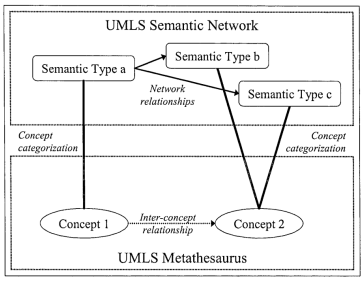

Figure 4: Structure of UMLS (Bodenreider and Burgun, 2005)

For research purpose, all the vocabularies provided by UMLS are available under the agreement of license. UMLS also supports a GUI where a user can enter a string or CUI (Concept Unique Identifier) or SNOMED code and retrieve information about it.

4.2 SNOMED CT, RXNorm, MeSH

SNOMED CT (Stearns et al., 2001) provides clinical healthcare terminology and it is highly comprehensive. SNOMED CT contains SNOMED code which uniquely identifies the clinical term which results into consistent representation of the term. It supports mapping to other ontologies. It is globally used in more than 80 countries. It is multilingual and also covers synonyms and definitions. There are 18 top level concepts in SNOMED CT such as body structure, finding, organism, physical object, etc.

In biomedical domain, drugs are the Named Entities. RxNorm (Liu et al., 2005)(Normalized names for clinical drugs) is the vocabulary which has assigned unique identifiers to drugs. This vocabulary is very useful in pharmaceuticals. It is also considered to be helpful for the exchange of drugs. Drug names are normalized by RxNorm. Doses of drugs, ingredients and its strength is also mentioned by creating relationships to the drug instance. Indexing of books and articles for life sciences is provided by Medical Subject Heading(MeSH) (Lipscomb, 2000) Dataset.

4.3 Anatomical Ontologies

Anatomical sites of the body parts can be viewed in hierarchical way. Brain can be divided into intratentorial region and supratentorial region, further supratentorial region contains basal cistern. This hierarchy can be converted to ontology and different relationships between anatomies can be incorporated.

4.3.1 Foundation Model of Anatomy

FMA (Rosse and Mejino Jr, 2003) is a public anatomical ontology, which is used as anatomical part of the UMLS metathesaurus. Structure of FMA makes it convenient to use it as a reference ontology and link any additional hierarchy on top of it. FMA has been made using a tool called protege. Classes defined in FMA can be seen in the figure 5

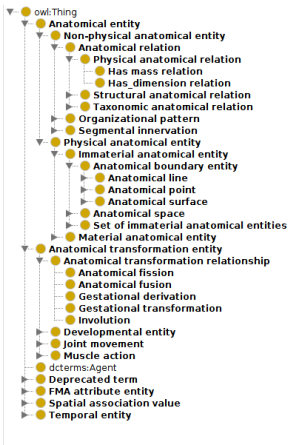

Figure 5: Classes of Foundation Model of Anatomy

4.3.2 SNOMED Anatomy

It is a substructure of SNOMED Clinical Terms which contains Anatomy class. SNOMED anatomy contains 26729 classes. As it is part of SNOMED CT, many ontologies which normalize the anatomy, also provide SNOMED code(unique identifier) for the same. SNOMED CT is widely applicable and so is the SNOMED code. This helps to link other resources of anatomy to the SNOMED Anatomy.

5 Ontology Learning and Knowledge Graph Creation

Ontology refers to an organized collection of concepts and classes in a particular domain area and how they are linked to each other. An ontology can be thought of as a skeleton of Knowledge Graph, i.e. , an Ontology once populated with the data at hand becomes a knowledge graph.

(Asim et al., 2018) in their survey have discussed methods of ontology learning. They have divided ontology learning techniques into three classes, namely statistical, linguistic and logical. Statistical analysis includes techniques like clustering, rule mining, co-occurrence analysis, etc. Linguistic techniques include Part of Speech tagging, parsing, lemmatization, dependency analysis, etc. They have also discussed evaluation techniques for ontology. Similarly for knowledge base construction, the techniques have been classified into four classes.

According to (Nickel et al., 2016), following are the classes under which techniques of knowledge base creation fall:

- Curated: Closed group of experts manually create triples.
- Collaborative: Open collaborative group manually create triples.
- Automated Semi-Structured: learned rules,regular expressions or handcrafted rules are used to extract triples from semistructured data like wikiboxes.

- Automated Unstructured: Machine learning approaches for extracting triples from unstructured text.

In Biomedical domain CTAKES (Savova et al., 2010) is widely used for entity recognition which can be helpful to create Knowledge Graph. It is a Natural Language Processing tool. It is a part of UIMA(Unstructured Information Management Architecture) and can be used to extract medical information from the electronic medical reports.

(Ogbuji, 2011) discuss about an ontology for computerized patient record maintenance. Their ontology is concerned with the entire procedure a patient has to go through while being admitted to a hospital including medical history screening, laboratory tests, clinical findings etc.

(Monteiro et al., 2016) in their paper mentioned creation of knowledge base from radiology reports. They provided pipeline for extracting and summarizing information from radiology reports into an ontology model.

6 Conclusion

In this paper, we have discussed the concept of knowledge graph and briefly described its components - entity and relation along with its applications and challenges in creating one. We, then, studied some of the existing knowledge graphs and their specifics. Further, we moved on to study some of the ontologies from the domain of our interest.

We also have covered literature that studies various techniques of creation of an ontology and the methods of creating a knowledge base.

References

Muhammad Nabeel Asim, Muhammad Wasim, Muhammad Usman Ghani Khan, Waqar Mahmood, and Hafiza Mahnoor Abbasi. 2018. A survey of ontology learning techniques and applications. *Database*, 2018.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Olivier Bodenreider and Anita Burgun. 2005. Biomedical ontologies. In *Medical Informatics*, pages 211–236. Springer.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan, and Stuart Nelson. 2005. Rxnorm: prescription for electronic drug information exchange. *IT professional*, 7(5):17–23.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Eriksson Monteiro, Pedro Sernadela, Sérgio Matos, Carlos Costa, and José Luís Oliveira. 2016. Semantic knowledge base construction from radiology reports. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*, pages 345–352. SCITEPRESS-Science and Technology Publications, Lda.

M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

Chimezie Ogbuji. 2011. A framework ontology for computer-based patient record systems. In *ICBO*.

Cornelius Rosse and José LV Mejino Jr. 2003. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of biomedical informatics*, 36(6):478–500.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. 2001. Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association.