Comprehensive Study of Semantic Annotation: Variant and Praxis

Sumit Sharma, Sarika Jain

Department of Computer Applications, National Institute of Technology, Kurukshetra, Haryana, India

Abstract

The proliferation of web content on the Internet has increased the demand for efficient information retrieval independent of content. The concept of the semantic web has revolutionized the way of searching, analyzing, and storage. Besides, semantic annotations provide esteemed solutions to enrich target information. There is a large amount of research available in the area of semantic annotations, which highlights the significance of annotation (such as sharing, integration, creation, and reuse, so forth) in various domains using annotation tools, be that as it may, none of these tools gives the earlier practice of the annotation research questions. Besides, no unified system exists that combines all the different kinds of annotations. This work presents a way to address the research questions given in the paper. We have combined isoforms of various types of annotations which have not been done to our knowledge till now. Furthermore, we have highlighted some prominent semantic annotation tools with their real-life applications, which depend on the type of annotation we classify.

Keywords

Semantic Annotation, Challenges, Applications, Ontology

1. Introduction

Information sharing and searching is a more useful task for Internet users but they are facing difficulties due to different representations of different data sources. Semantic annotation modeling can fill this gap of various knowledge representations. It establishes the relationship between the data entities and joins the term or mentions to entities. The objective of the semantic annotation measure is to survey what parts of the report compare the ideas portrayed in the ontology, and along these lines, the outcome is a bunch of mappings between record sections and ontology concepts as defined in [1]. Natural language technologies are one of the emerging trends of their use for the sciences and humanities. Experts are facing problems such as the explosion of information due to the continuous increase in the production of scientific content on the web, which makes it difficult to observe the state of the art in a given domain [2]. Semantic annotation applications have been used in different domains in different ways, but all of these have a common goal. Authors have applied the semantic annotation for the Arabic web document by deep learning methods [3]. Annotations can also contribute to manage natural history collections using semantic annotation [4]. Authors applied semantic annotation on digital music to improve the trend of searching music [5]. Thus, annotation can be termed as to reduce the mental effort when a document is read for the purpose of research and analysis. Therefore, the process of embedding additional information to the already available information helps to interpret the information, remembering things, traceability, machine understanding capability, and many more.

Another assumption about semantic annotation is to use a machine to understand the relationship between the URI and the network of data. If the text is semantically marked, then it becomes a source of learning which is easy to understand, consolidate and reuse by machines. Semantic annotation helps machines to use data on the web to self-interpret, combine results, and manage digital information from information available on the internet. Such information can be generated by interpreting sources from metadata that can result in "annotations" about all resources. In this paper, we shall examine semantic annotation by defining the annotation and metadata, and then we shall discuss various aspects of semantic annotation approaches and review the current generation of semantic annotation systems.

Here in this paper, we are preparing and addressing some research questions, which are benignant and significant for the research development of meanings and annotations. We have described isoforms of various kinds of annotations with a formal description of the semantic annotation to a nexus between research questions. We are also going to explain essential as-

ACI'21: Workshop on Advances in Computational Intelligence at ISIC 2021, February 25-27, 2021, Delhi, India

sharma24h@gmail.com (S. Sharma); jasarika@nitkkr.ac.in (S. Jain)

https://sites.google.com/view/nitkkrsarikajain/home (S. Jain)

0000-0001-5054-8670 (S. Sharma); 0000-0002-7432-8506 (S. Jain)

pects of semantic annotations that are being used for diversity of semantic annotations related to different domains. Furthermore, we have highlighted some exigent semantic annotation tools alongside their real-life applications, which depend on the type of annotation we classify.

2. Research Questions

Here in this section, we provide a brief study on semantic annotation to elaborate major research questions "what, where, why, and how" to use the semantic annotation. "What?", describe the definition of annotation, "Where?", examine where to apply, "Why?", define the importance of annotation and "How?", define various ways to represent annotation.

2.1. What? (Definition)

According to Oxford Dictionary Online, the sound "annotation" is defined as "a note by way of explanation or comment added to a text or diagram"[6]. Semantic annotation contributes to mark-up the existing texts to justify their senses so that a machine can automatically identify and process information, thus making them more valuable. In literature, the definitions as employed by different authors for semantic annotations were quite different. In any case, Semantic Web achievement relies upon the accomplishment of an extraordinary number of clients semantic substance. This accomplishment requires apparatuses that decrease the multifaceted nature of semantic innovations. Semantic annotation is the fitting procedure for searching a word, sentence, and paragraph semantically in the Semantic Web. Annotations are also used to transform syntactic structures into knowledge structures.

All the more succinctly, annotation or tagging is a process that allows to draft a section, statement, comment, or attributes to a document or segment in a report. When all is done, the annotation can be viewed as additional data related to a specific point in one record or another snippet of data [7]. The authors [8] present an overall meaning of annotation as including some other bit of information and further expanding the definition of annotation in various domains. Generally, domain annotations are typically labeling a concept (record, part of an archive, or word) legitimately to perceive the essential concept or principle thought in the information. The tagging helps users to recognize or classify a document based on the concepts required and also helps to target the outcome of the document [7].

Semantic annotations represent transitional formulation of connections between unstructured documents, semi-structured documents, and ontologies in both directions [9]. Embedding metadata with the documents to assign semantics on the web assets is a semantic annotation by innovative judgment [10]. All the above definitions provided by various authors have one thing in common: linking resources with domain ontology.

2.2. Why? (Purpose)

This research question is the most important research question to solve the significance of the development of annotation. The textual data's growing phenomenon requires Natural language processing and text mining procedures to arrange and recognize patterns and knowledge from the texts. The need for semantic annotation is becoming important because the information is represented as a knowledge graph [1]. Data is regularly traded in an electronic arrangement (like papers, letters, note amalgamation, mail, data set, report, laws, proposals, articles, and declarations). The purpose of semantic annotation encourages the semantic web-enabled machines to self-interpret, consolidate the results, and practice it on the web. We can create such information by annotating sources using metadata, outcoming in "annotations" concerning that source.

Probing, searching, mining, and classifying are growing significant and challenging jobs with extensive massive data. This job grows even more complicated if the data explode, and the data are undefined. It is not straightforward to manually read all the documents and find a particular concept (person, event, place, so on) in the full document. Annotation provides a significant role in the search for any key idea in the documents. It is challenging and essential to discover all the key concepts and relationships in the documents during annotation. Exploring the relationship between data concepts and rendering it in a new form is again the discovery subject. Ontology is the right way to define the relationship between data concepts, and simultaneously, it provides advantages to data in the form of machine understanding.

2.3. Where? (Place)

In the last few decades, the experimental form of annotation has grown a lot. We have found the usability of annotation in various fields. There are extraordinary implications and uses in various areas of annotation. Programming languages use annotation on class, method, parameters, or variables for their clarity and definition. On the other hand, mechanical en-

gineering uses annotations to understand the specific meanings of text or symbols. Before the utilization of annotation, it is valuable to think about the scope of annotation that exists so that anyone can pick the correct type for their use case. Annotations incorporate a broad scope of data types on which it tends to be applied and reuse. Some essential ranges of annotation include text, image, audio, video, graphics [2]. Some annotation tools have evolved to show the use cases of annotations that provide a lightweight framework to annotate textual data [11]. The authors [12, 13, 14] applied semantic annotation on image data to improve the searching. Likewise,[15] provided a methodology to add an annotation to XML schemas same as[5] apply annotation on digital music. Semantic annotations are useful for digital document classification (newspapers, blogs, media content filtering). It is possible to search for a particular concept (named entity recognition) in large amounts of data. Annotation has an essential role in the biomedical field to identify essential terms used in medicine [16]. Currently, IoT sensor data are being stored by meaningful annotation to clearly express the powerful potential and impact of the data [17, 18].

2.4. How? (Implement)

This is the most important research question that plays an imperative role in the success of annotations. Also, the applicability of the semantic annotation depends on the nature of the data type. It can be text, image, audio, video. For the text annotation, it could be Semantic Annotation, Intent Annotation, and Named Entity Annotation. Finding the essential concept in the text is the main work for a text document. Image annotation is essential for an extensive scope of utilizations, including PC vision, automated vision, facial acknowledgment, and arrangements that depend on AI to decipher pictures. To prepare these arrangements, metadata should be doled out to the pictures as identifiers, inscriptions, or catchphrases.

In the last few decades, several techniques were developed for semantic annotation. The part of speech (POS) annotations depends on the specific design and model demanded. One may be interested in a limited POS annotation scheme if one wishes to do text mining or text processing. Semantic comment stages offer help for data extraction advancement, knowledgebase and ontology executives, warehouse, access APIs (e.g., RDF repositories), and UIs for knowledgebase editors and ontology [19]. The semantic annotation is likewise helpful for a legitimate grouping of e-reports, online news, web journals, messages, and computerized

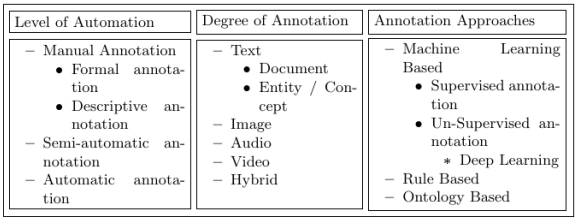

Figure 1: Various aspects of Semantic Annotation

libraries that need text mining[20], AI, and natural language handling techniques to get meaningful information. As per our knowledge, the most popular machine understandable format nowadays is RDF (W3C, Resource Description Framework (RDF) http://www.w3.org/RDF/. Last accessed January 25, 2021.).

3. Preliminaries for Semantic Annotation

In our studies, various aspects of semantic annotations are shown in Figure 1, which completes the survey of semantic annotations. This section is important to know the structure of annotation, here we shall begin with the basics to describe the complete method of practicing annotation. Then, based on the structure of data types in which semantic annotations addressing the research question and then provide a formal definition of semantic annotations to serve the purpose of annotations.

3.1. Semantic Annotation

Annotation is the process of allocating some labels to the data for data interpretation and automatic description. Semantic annotation is the annotation in which some necessary additional information is added to a text document to reflect the relationship between ontology class concepts or instances and text document entities. This brief description of the object defined consists of the main body of the paper. It describes semantic for a document (such as label, title, author, date of publication, etc.). Therefore, semantic annotation collects semantic information from intuitive and more essential records so that target information can be easily searched and classified by the machine.

The annotation output of a document can be in different forms and depends upon the tools or methods that produce annotation. The goal of the annotation project may differ according to the design and requirement of the project model. Figure 2 shows an example

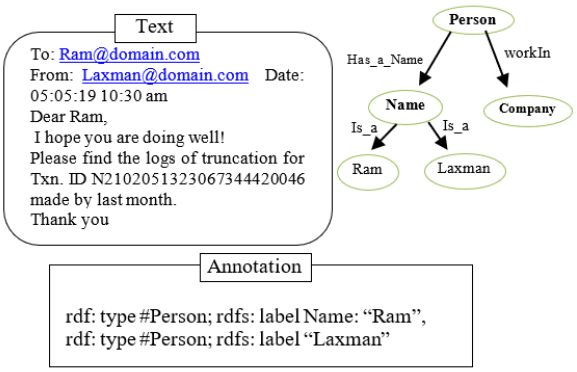

Figure 2: Example of semantic annotation

annotation of email text data annotated by ontology.

3.2. Types of data

In the present scenario, annotation is one of the most challenging tasks as data on the web is not uniform (different structures). Semantic annotations can be applied keeping in mind the nature of the data. Therefore, it is essential to provide a unique description of the data to make the data different and to avoid annotation problems. Motivated from this, in this section, we will throw light on various types of data (on the internet) which is significant to semantic annotation. There are three kinds of data namely; structured, unstructured, and semi-structured, which are explained below.

3.2.1. Structured Data

A well-Organized form of data is known as structured data, which is easy to explore and generally arranged in rows and columns (e.g., excel spreadsheet). In the structured data, a portion of the information always periodically outlines into fixed predefined attributes, which is occurred in form of the columns. For instance, Table 1 shows the structured form of the transaction log in which the excel spreadsheet, database designer designs a data model that is followed to store the structure data. This is the best example of structured data. This data model saves all records into a table. These records are collected with the help of a relationship that exists between the entities of data. Structured data utilize the storage space and make information retrieval as easy as possible.

SQL, MySQL, and SPARQL are the query languages used to retrieve, manipulating, and storing the structure data. These query language groups the database

Table 1
Example of structured data

S.N	Account No	Name	Transaction	Logs
1.	XXX445050	Ram	4393949	5:5:19
2.	PPP304039	Laxman	2932734	3:4:19

To: Ram@domain.com
From: Laxman@domain.com Date: 05:05:19 10:30 am
Dear Ram,
I hope you are doing well!
Please find the logs of truncation for Txn. ID N21020513230673444200 46 made by last month.
Thank you

Figure 3: Example of unstructured email message data

on the relationship defined in the data model. Structured data can be handled by humans as well as by machine. However, human has less role in the annotation and structured data are easy to annotate by some predefined rule [21]. These rules are created based on the relationship between the entities.

3.2.2. Unstructured Data

Several authors have worked on the other form of unstructured data like (images, audio, video, news, social media data, blogs, open-ended survey, web content, transcripts, etc.). Various AI and Machine learning-based algorithms have been applied to recognize the content and then annotate them accordingly. It also provides a hidden association between the entities with the help of links. The wide range of data on the Internet is unstructured data. Generally, heterogeneous data cannot be stored as a row and column and does not have an associated data model. The general examples of unstructured data are web email, blogs, and HTML pages. Since there is no underlying relationship between the data concepts, therefore, finding, analyzing, accessing, and managing a piece of information in this kind of data is more complicated. According to some machine learning algorithms [22, 23, 24], these processes are erroneous and time-consuming tasks. Figure 3 shows an example of unstructured data.

3.2.3. Semi-structured Data

Semi-structured data is another variety of data that mix the structured and unstructured data. It has re-

Figure 4: Example of semi-structured data on web search of paper

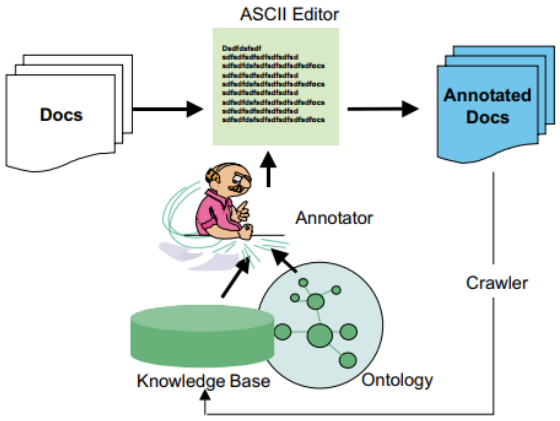

Figure 5: Manual annotation of document with semantic data

markable properties to organized information but does not relate to the fixed structure of the data model. Web forums, web pages, and email messages are the popular examples of semi-structured data in which, the actual content is unstructured, and this form of data also contain some structured information such as name and title, log information, time, etc.

Figure 4 shows an example of semi-structured data about a web page. That also contains some structured information about the web page like title, journal name, journal log, etc. This semi-structured data provides a little help to the designer to build the data model. These small pieces of information involve extracting data from the unstructured repositories.

3.3. Level of Automation of Annotation

Successful use of the Semantic Web requires far reaching accessibility of semantic annotations for existing and new records on the Web. The level of automation shows how we can get the right data and how to use it correctly. It defines the automaticity of the machine from manual to automatic. The level of automation in any systems can be assessed, measured as manual, automatic, and semi automatic described in [7, 9, 25] with their framework and requirements.

3.3.1. Manual Annotation:

Manual annotation is a process of reading an input document and extracting a piece of new information with human participation. Manual annotation is also a type of formal annotation with human computer interaction. It tracks many NLP tasks and has lots of activities [26] such as writing comprehensive annotation guidelines and defining an annotation schema, etc. Manual annotation is even more conveniently developed today, utilizing writing tools, such as Semantic Word [27], which give an incorporated atmosphere to authoring and annotating text. Notwithstanding, human annotators' utilization is as often as possible due to components, for example, annotator knowledge of the domain, a measured amount of training, personal inspiration, and complex patterns. Manual annotation cannot be applied to a massive portion of data. The semantic annotation of archives concerning an ontology and an entity knowledge base is examined in [15]. Even though introducing intriguing and yearning draws near, these do not talk about the utilization of robotic strategies. The center is the manual semantic annotation for the enrichment of web content, while few cutting-edge manual annotation approaches are examined regarding difficulties of supporting multiple formats (HTML toward PDF, XML, images (e.g., PNG, JPEG), and video. For a depiction of some more established tools or frameworks, please allude [28]. The authors also provide a classification of semantic annotation system detailed analysis of end-user tools, pros, and their cons.

The manual annotation tools allow humans to add some description of text to web contents or the other sources of data. However manual annotation has become very complicated because of its usability and feature [29]. Protégé [29], SMOR[30]E, and OntoMat [31]. The author[26], have provided the list of annotation tools based on the detailed evaluation of annotation feature Besides this, manual annotation is time con-

suming and often full of errors. As shown in the Figure 5, it requires expert knowledge for being domain-specific. For manual annotation, a large volume of training is needed. Due to the complex schemas, it is also not easy to handle large-scale data, and there is no reuse of output data. Human annotation is too costly and time consuming and cannot be applied to control the massive amount of records available on the Web. Manual annotation requires qualified annotators, this has been explained with the help of an example in section 4, and first, an annotator would map the text "Ram" to domain ontology and recognize it as a Person and further would recognize the company, where Ram is working. Based on tagging of the data, manual annotation is further categorized as formal and descriptive annotations.

- **Formal Annotation**
  Formal annotation is the simplest and fastest way to annotate documents by the human. In the formal annotation, some scripts are added to the record such as (title, author, publishing date, etc.). To do such a task, experts do not require detailed knowledge about the domain, only conceptual understanding is needed.

- **Descriptive Annotation**
  A descriptive annotation or summative annotation can describe the main goal of the work. Descriptive annotation provides a summary as well as a complete citation of the job without evaluating the quality of work. Descriptive annotations include an overall description of objects that may be enough for the machine to understand the full semantics of the material and process the information. For example, it means to convey a book, hypothesis, methodology, article, conclusion, or any other source.

3.3.2. Semi-automatic annotation

In a semi-automatic semantic annotation, the framework creates an annotation and these few are then post-edited and amended by human annotators [32]. Many manual annotation tools transferred to the semi-automatic framework by providing manual training. Researches on semantic annotation methods investigate the benefits of a state-of-the-art tools for semi-automatic to help the semantic annotation of a large set of biomedical queries [16]. There are numerous semi automatic semantic frameworks, MnM [33]. Unlike manual and automatic ones, don't consolidate programmed into the semantic investigation, however, either use them as an extension between models elements and annotation. Semi-automatic annotation requires a mixed structure in the annotation model that has increased the structure complexity [25]. This kind of annotation model is fit for supporting labels or tags that are not related to a specific property but on the other hand are portrayed to depict a particular connection among metadata assets for navigation purposes seen at [9].

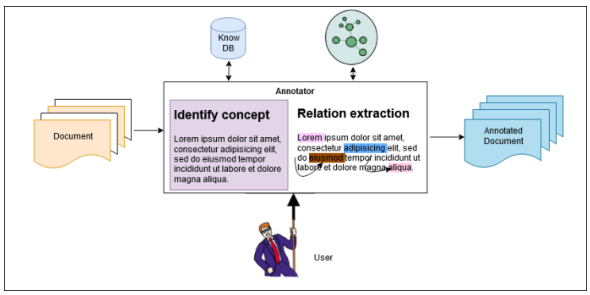

Figure 6: Semi-automatic annotation process of documents with semantic data

The semi-automatic annotation is shown in Figure 6, in which both human and machine become the annotators. Semi-automatic is fast and robust to find the semantic relationship between the annotating data and the targeted annotated document. Human enrollment provides a significant advantage to semi-automatic annotation to adopt the new feature and new domain. Morphological analysis, part-of-speech tagging, retrieval of domain-specific information, and recognition of name entities are the significant component of semi-automatic annotation.

3.3.3. Automatic Annotation

Automatic annotation is a high level of semantic annotation. Systems falling into this category are highly trained and have high accuracy. To train this type of system, a large amount of quality data and rule sets are required. To deal with these issues, unsupervised systems tried the many methodologies and experiment to learn how to annotate data without human oversight, but precision is as yet restricted. The automatic meaning of lexical data allows both annotations to add important information to the production search and index the document [16]. Article [26, 24] proposed a scientific classification for information extraction tools dependent on the principle strategy adopted on a larger scale by the community. Some other techniques use machine learning methods [22] to automate the semantic annotation using some training data.

Automatic semantic annotation is controlled by a machine, so this annotation is efficient and is fast as

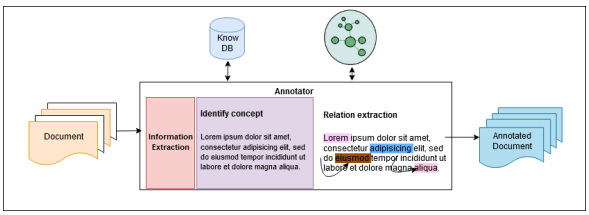

Figure 7: Automatic annotation process of documents with semantic data

compared to manual annotation. The main key feature of an automatic semantic annotation is that it can handle massive data, which is the limitation of manual annotation. In automatic annotation, absolute rule or standard schema must be defined to work machines efficiently. Based on fascinating predefined standards, the automatic annotation performs the task. Automatic annotation is useful for dynamic web content that may be transient. Automatic annotation entirely depends upon the training module and failed to adopt new terminology. However, the complete automatic semantic annotation for global data still is an unsolved problem. Hence, semi-automatic annotation methods are being used widely in current scenarios. The component of the automatic semantic annotation is shown in Figure 7, in which no interaction of humans at the running state.

3.4. Degree of Semantic Annotation

With the development of semantic annotation in the most recent couple of years, the semantic annotations can be applied in various spaces to extend convenience. In the absence of the structure of web data, automatic discovery of targeted or unexpected knowledge actually develops various research issues outlined in [22]. Heterogeneous data could be text, picture, sound, video, illustrations. The authors in [2] applied semantic annotation textual objects and provide the practical impact of semantic annotation on the search. And in [12] applied semantic annotation on image objects to improve the searching and indexing. On the other hand, [15] gave a procedure to add an explanation to XML compositions. To the best of our knowledge, no such annotation technique exists that can be successfully applied to all content (text, image, audio, video) simultaneously. To keep in mind that diverse strategies are used for different content, we can be classifying the annotation as a degree of annotation to use the common framework of the semantic web. Semantic annotators take input in a variety of forms, which is known as the degree of semantic annotation. It makes the system flexible. The degree of annotation defines the classification of annotation based on the input structure of data as shown in Figure 1.

3.4.1. Text:

Most of the information on the web is in the form of text. Data is extracted from the web directory through a user query. This user query is also written in form of text only. The input query could be mapped on structured, semi-structured, or unstructured data. Annotation of a text document is important to search, analyze, and classify the documents correctly.

3.4.2. Image:

Increasing digital capturing techniques have led to a fantastic evaluation of images on the web. A text query is used to access a huge amount of image data sources. To achieve this, a query is written which produces a visually similar description to the image. This feature of the image becomes a key to represent it. Several annotation techniques have been used to make and describe the main feature of the image. Some of the researchers have focused only on the feature extraction method and have developed an image semantic annotation method based on an image concept distribution model.

3.4.3. Audio:

Universal mobile interface makes digital cities portable with audio Earth annotations. The best example is cities carriable with audio semantic annotation and that aimed to provide a comprehensive mobile interface to the mobile user on demand.

3.4.4. Video:

Video annotations are equally important as image and audio on the web. Video lectures, social media content, news video, sports, etc. are the data that is monitored by semantic annotation. In the semantic context of the examined domain, the concept, instances, with their visual descriptors, enrich the video semantic annotation.

3.4.5. Hybrid:

Multimedia content base semantic annotation is more challenging and based on high-level ontologies. These approaches are on demand.

4. Approaches for Semantic Annotation

Several approaches have been proposed to explain the need for semantic annotations in user information and vast knowledge spaces. Some of the strategies focused on semantic tagging (i.e., title, author, explanations, etc.) in the document to annotate. It has reduced large volume of search that need to find supplementary information in external sources [34, 35]. It categorizes the annotation tools according to the media content, which can be annotated by annotators (for example, text, audio, video, images, etc.). Furthermore, in viewpoint to know how to achieve semantic annotation, there are various approaches and techniques used to achieve annotation. [22] investigate the machine learning approach to automate the annotation process. Automatic semantic annotation is more effectively finished nowadays, utilizing machine learning techniques. We can further categorize the semantic annotation into various automatic approaches, including Supervised machine learning based method, Unsupervised machine learning based method, Rule based methods, and Ontology based Machine Learning. Supervised approach is completed in two stages, training and annotation. In the training provide the plain text with some labeled and in the annotation, the machine has to recognized entity and semantic relation based on the training labeled data. [12] apply supervised machine learning techniques to annotate image data. In an unsupervised approach, make an annotation with unlabeled data. For instance, [12] proposed a strategy for automatically summing up the extraction designs from the website pages. The ontological annotation approach utilizes other information sources like Wikipedia, Vocabulary, thesaurus ontology, etc. Rule-based semantic annotation is based on some pre-defined rules. Rule-based algorithms for semantic annotation, various extraction frameworks have been created based on the strategy, for instance: Crystal [36], AutoSlog [37], MnM [33], Rapier [38], SRV [39], Whisk [40], Stalker [41], and BWI[42]. The rule-based approach [43] is only applicable if the streaming pattern is well known. It is difficult to apply to the heterogeneous unknown structure.

4.1. Machine Learning Methods

The dynamic environment and a wide range of domain influence the system to perform automatic annotation. The automatic annotation process is one of the critical and challenging tasks for a semantic annotation system.

4.1.1. Supervised Machine Learning

In a supervised learning method, an expert assigns the key to annotating data. To deal with annotating data, several supervised machine learning models such as SVM, Hidden Markov Model (HMM), Markov Random Field Model to be implemented to optimize labeling costs. In this approach, firstly a pair of entities are mapped with the web as a corpus, then it finds a binary relationship between the entities and if a relation is found, then it labels as a favorable otherwise marked as unfavorable.

Furthermore, some authors have extended an existing approach with the help of the SVM machine learning technique but the main drawback of this method is that it cannot handle the multiple instances of learning and during process, many bugs are found. Other challenges in semi-supervised and unsupervised techniques to retrieve relation between the entities are discussed in [44].

• Limitations of supervised machine learning approaches

- Large Training Corpus: The efficient machine learning model requires significant expert annotated corpus for training purpose and which are very expensive to develop.
- Limited Entities Extraction: This machine learning models have only identified entities on which models were trained. Other remaining categories of entities which are not recognized generate a false result, which affects the accuracy of the model.
- Lack of entity relation: Due to large data corpus, it only explores the surface of the graph for every instance of knowledgebase.

Supervised machine learning methods are expensive and require a lot of effort. So, most of the research has moved towards unsupervised or semi-supervised machine learning methods. These methods have been discussed in the next section.

4.1.2. Unsupervised Machine Learning

Unsupervised machine learning is the process of automatically identifying possible relationships between objects of massive text corpora. Unsupervised machine learning methods do not require manually labeled data. Pairing deep learning with unsupervised learning crosses

the boundaries of supervised learning. This machine learning method clusters similar entities concepts. These clusters are commonly used to describe relationships of sets that occur in such a way that the elements of sets refer to the same group. Researches examine some clustering techniques with some of the novel approaches discussed in [45]. They have created a simplified and generalized grammatical clause representation that utilizes information-based clustering and inter-sentence dependencies to extract high-level semantic relations. [46] discovered and enhanced concept specific relations other than global connections by web mining.

• Limitations of unsupervised machine learning technique

- Due to automatic nature, sometimes it generates unnecessary clusters that were not an area of interest.
- The output is less accurate because one input data is not known, and the data expert does not label dynamically.
- It does not extract the hidden relationship between the entities and does not provide the link to relation.

4.1.3. Deep Learning Method

Due to large interlinked datasets on the internet, machine learning aims to provide a method that processes data automatically. The idea could be achieved in the present text using deep learning semantic annotation based on public and common ontologies. Due to the gradual growth and the large size of the resources, there is a need to have an active and quick semantic annotation of resources. For example, Neural Network, CBOW, and Skip-gram have become the state-of-the-art for generating word embedding. The authors [21] have presented a deep learning and rule-based learning technique for the Arabic language which involves discovering a document and used to enhance the semantic indexing.

4.2. Rule-Based Annotation Methods

Rule-based annotation is the simplest and most straightforward approach, which depends upon a predefined rule created by one or more experts. The rule base annotation can be applied only when either the data is fully known or have some specific notation. For example, the rule base annotation is perfect for structured datasets such as RDBMS data. Rule base annotations cannot be applied to other types of data or unstructured data. In this type of annotations, experts write some rules with the help of logical arguments, so that the relationship can be extracted by carefully observing the correct logic. The rules follow some specific IF-THEN-ELSE formats that elicit information from a high-level reference using a low-level reference. According to our survey of the literature, rules have been applied when it combines ontological reasoning [21]. Author [47] have provided a minimal rule engine, MiRE, for a context-aware mobile device. The rule is significant and can be applied in various tasks like event detection, IoT data representation.

• Limitations of rule based approach

- It is applicable only to recognize regular pattern.
- Dynamic changes cannot be easily handled by this approach.
- Need expert to generate a rule with complete domain knowledge.
- Need large and complex rule to deal with unknown vast data set.

4.3. Ontology-Based Methods

Ontology-based, dictionary-based, or knowledge-based semantic annotation is the most robust annotation approach to represent a relationship between data objects. Ontology-based semantic annotations can be applied with any automation category (manual, semi-automatic and automatic annotations). As we have discussed in Section 5, this annotation approach introduces the process of generating metadata using ontology as their knowledge base. The ontology-based approach relies entirely on description logic, which relates to a family of logic-based knowledge representations of formalism. All ontological reasoning approaches have been supported by two general illustrations of semantic web languages. i.e., RDF (S) [48] and OWL [49, 50].

Several frameworks support manual annotation, for example, Protégé-2000, CREAM , SMORE, Artequakt are the semantic annotation framework that supports various semantic annotation task (like create an annotation, add a tag, validate, etc.). Knowledgebase tools help to manage and store complex information. Some annotation tools have been used to develop and maintain the dictionary of the document. ERASMUS and SIBM (CISMeF), NCBO Annotator, are some concepts Mapper used to map the concept of a word to the instance of the dictionary.

Many semantic query languages (such as Triple, RQL, SPARQL, RDQL, etc.) and various reasoning engines (RACER, Pellet, and FACT, etc.) connect the semantic web languages. Some techniques such as the SWRL rule provide popularity to ontological reasoning. Ontological modeling represents the knowledge in a hierarchical form and establishes the link between the related entities.

- **Limitations of Ontological approach**
  - Ontological modeling is domain specific.
  - Expert knowledge is required to genereate a query.
  - Ontology-based query engine required to retrieve information.

5. Semantic Annotation Tools

We can arrange annotation tools in a two-dimensional space, Ontology Support Semantic Annotation tools and Non-Ontology Support Semantic Annotation tools. Describing these tools based on the various aspects of semantic annotation.

5.1. Non-Ontology Support Semantic Annotation Tools

We are highlighting the most frequently referred non-ontology-based tools found in the literature study of current semantic annotation. These tools annotate manually and some use different strategies to reduce the effort of annotating. Some tools have the option to perform annotation manually as well as automatically and some have option both (semi-automatically). Some important semantic annotation tools are shown in Table 2.

5.2. Ontology Support Semantic Annotation Tools

Current semantic annotations, based on the literature, aim to support the development of inter language resources. Many researchers are working in this area and several authors have contributed in multiple ways to make it successful. They have defined semantic annotations in a different appearance but have the same semantics. Some ontology-based semantic annotation tools and their aspects are shown in Table 3.

6. Advantages and Applications of Semantic Annotation

The advantages of annotation include searching, storing, analyzing, and automation. In this section, we shall discuss the various benefits of semantic annotation and its real-life application.

6.1. Benefits of semantic annotation

The semantic annotation helps to formulate logic for a more profound understanding by the machine. Semantic annotation is encouraging the researcher to make inferences and draw conclusions about web resources. Some of the benefits of semantic annotation are given below.

6.1.1. Improves searching:

Searching the vast and distributed structure of the web requires efficient search schemes. Searching becomes efficient when the available information is meaningful and contains meta-data to support the information available on the internet. The semantic search will be defined as a search that is based on semantics rather than just depending on text similarity[51]. Semantic annotations are also used to correlate significant tags among reports to perform a semantic search.

6.1.2. Better utilizes the available web resources:

Now a days, when almost everything is well defined, organized, and adequately classified on the Web, then the resources can be efficiently utilized. The information is available on the Web in various forms such as document, knowledge base and dictionary, etc. contains information in the form of text or image or both can be linked appropriately through annotation. The semantic annotations of web resources are connected concepts with meaningful representation in which the retrieved information could be utilized according to user interest instead of just a text matching.

6.1.3. Improves the decision making:

It has been found that when all the related and significant data has been shared with the clients (through semantic look), at that point, the client is capable of making a few choices and can perform it successfully since he/she will be mindful of all the things. The semantic search will be helped by semantic annotation

Table 2
Non-Ontology Support Semantic Annotation Tools

SA System	Approach	Description	Application domain	Automation
BroMo	Unsupervised	Using clustering for blogs and article semantic annotation	Proteins (biomedical)	Semi-automatic
Sozekamm	Supervised	Annotate data using a supervised categorical clustering algorithm LIMBO	Gemeral	Semi-Automatic
Ontea	Unsupervised	Process email or text document find the pattern	Text and Email	Manual/ Automatic
Doccano	Unsupervised	Open source text annotation tool for human	General	Manual
Yawas	Unsupervised	Java based web-based annotation system	General	Manual
Briefing Associate	Unsupervised	Used for Microsoft power point presentation	MS Power point	Manual
Zemanta	Rule based	Algorithm for natural language and semantic processing is proprietary	General	Semi-Automatic
Thresher	Unsupervised	Aimed to Web pages with similar content	Web page	Automatic/Manual
RCSSAT	Supervised	Classify the using a new lexicon	General	Manual

since the semantic search is concerned with the meaning of the substance accessible. The semantic annotator has given a semantic search with proper explanations to empower it to make an appropriate sense in the document, picture, etc.

6.1.4. Unambiguous description of abbreviations:

Many words/concepts have been expressed using the same abbreviation. This leads to a critical problem of ambiguity. The use of annotation is an effective way to troubleshoot this problem.

6.1.5. Automatically classifies the web resources:

If the resources available on the web are annotated properly, then the classification process will be uninterrupted because all classification algorithms only ask for the references of annotated metadata to classify the resources. This makes semantic web search efficient as a process of classification.

6.2. Applications of semantic annotation

After specifying the structure model of semantic annotations, annotation creators can apply the annotation to serve their purpose such as (search, sharing, integration, reuse, etc.). Here, in this section, several applications of semantic annotation are listed with some real-life applications.

6.2.1. Bibliographies:

Semantic annotation plays a vital role in the field of bibliography annotation to describe the source. The whole information of the source is essential for the authors while writing a paper. Bibliography semantic annotation helps in linguistic data to analyze and is used for any language data.

6.2.2. Extraction of open information:

Semantic annotation has been practiced in diverse fields of knowledge. For instance, It has an application in a news analysis for the naming of places, organizations, and people. it has application in biological systems for

the identification of biomedical entities such as genes, proteins, and their relationships.

6.2.3. Alignment of ontologies:

This is one of the important applications for the alignment of ontologies for knowledge management. Ontology alignment is quite useful to differentiate the heterogeneous models and it relates the difference to determine various interoperability concerns that synchronize in semantic image annotation and retrieval.

6.2.4. Semantic search:

Search engines can retrieve the required documents more accurately with the help of metadata information. Scientists and librarians put lots of efforts and time to create metadata for the documents. However, to alleviate the hard labor, many attempts have been made towards generating the automatic metadata, based on the techniques of information extraction.

6.2.5. Classification:

Semantic annotation helps to classify the data which speeds up the task and makes data secure. The information retrieval-based system on semantic annotation helps to manage the data according to the search interest of the user.

7. Conclusion

The purpose of this paper is to distinguish integrated review methodology from other review methods and to propose research questions for integrated review methodology to increase the rigor of the process. In this paper, we have presented an extensive study of important approaches used for semantic annotation of a text document and a wide variety of approaches to explore the prominent historical semantic annotation models applicable for text document annotation. We have also provided the various aspects of semantic annotations based on which annotation can be classified. We have also highlighted the importance of semantic annotation in real-life practices.

The comparison of semantic annotation tools has been done through a level of automation, degree of annotation, and type of annotation. As a future scope, we are also trying to implement the semantic annotation models which will map with the global corpus and can be applied to any domain. Finally, this approach also has been compared relatively with other ones available in the literature.

References

[1] F. Pech, A. Martinez, H. Estrada, Y. Hernandez, Semantic annotation of unstructured documents using concepts similarity, Scientific Programming 2017 (2017).

[2] H. Agt, G. Bauhoff, R.-D. Kutsche, N. Milanovic, J. Widiker, Semantic annotation and conflict analysis for information system integration, Proceedings of the MDTPI at ECMFA 2010 (2010).

[3] S. Albukhitan, A. Alnazer, T. Helmy, Semantic annotation of arabic web documents using deep learning, Procedia computer science 130 (2018) 589–596.

[4] L. Stork, A. Weber, E. G. Miracle, F. Verbeek, A. Plaat, J. van den Herik, K. Wolstencroft, Semantic annotation of natural history collection, Journal of Web Semantics 59 (2019) 100462.

[5] F. Rahman, J. Siddiqi, Semantic annotation of digital music, Journal of Computer and System Sciences 78 (2012) 1219–1231.

[6] C. Soanes, Oxford dictionary of English, Oxford University Press, 2005.

[7] E. Oren, K. Möller, S. Scerri, S. Handschuh, M. Sintek, What are semantic annotations, Relatório técnico. DERI Galway 9 (2006) 62.

[8] V. Batanović, D. Bojić, Using part-of-speech tags as deep-syntax indicators in determining short-text semantic similarity, Computer Science and Information Systems 12 (2015) 1–31.

[9] K. Bontcheva, H. Cunningham, Semantic annotations and retrieval: Manual, semiautomatic, and automatic generation, in: Handbook of semantic web technologies, 2011.

[10] W.-f. WANG, L. ZHAO, Research and application of ontology on semantic web [j], Journal of Zaozhuang University 2 (2007).

[11] M. Kogalovskii, Semantic annotating of text documents: Basic concepts and taxonomic approach, Automatic Documentation and Mathematical Linguistics 52 (2018) 134–141.

[12] G. Carneiro, A. B. Chan, P. J. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE transactions on pattern analysis and machine intelligence 29 (2007) 394–410.

[13] S. Prasad, A. K. Lodhi, S. Jain, Helpi viz: A semantic image annotation and visualization platform for visually impaired, in: International Conference On Computational Vision and Bio Inspired Computing, Springer, 2018, pp. 881–888.

[14] S. Jain, S. Prasad, A. K. Lodhi, Semantic annotation of images with text and sound for visually

impaired, Journal of Open Source Developments 5 (2018) 20–27.

[15] H. N. Talantikite, D. Aissani, N. Boudjlida, Semantic annotations for web services discovery and composition, Computer Standards & Interfaces 31 (2009) 1108–1117.

[16] A. Névéol, R. I. Doğan, Z. Lu, Semi-automatic semantic annotation of pubmed queries: a study on quality, efficiency, satisfaction, Journal of biomedical informatics 44 (2011) 310–318.

[17] S. Balakrishna, M. Thirumaran, V. K. Solanki, Iot sensor data integration in healthcare using semantics and machine learning approaches, in: A Handbook of Internet of Things in Biomedical and Cyber Physical System, Springer, 2020, pp. 275–300.

[18] S. Jabbar, F. Ullah, S. Khalid, M. Khan, K. Han, Semantic interoperability in heterogeneous iot infrastructure for healthcare, Wireless Communications and Mobile Computing 2017 (2017).

[19] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, M. Goranov, Kim–semantic annotation platform, in: International Semantic Web Conference, Springer, 2003, pp. 834–849.

[20] N. Kiyavitskaya, N. Zeni, L. Mich, J. R. Cordy, J. Mylopoulos, Text mining through semi automatic semantic annotation, in: International Conference on Practical Aspects of Knowledge Management, Springer, 2006, pp. 143–154.

[21] C. Lhioui, A. Zouaghi, M. Zrigui, A rule-based semantic frame annotation of arabic speech turns for automatic dialogue analysis, Procedia Computer Science 117 (2017) 46–54.

[22] J. Tang, D. Zhang, L. Yao, Y. Li, Automatic semantic annotation using machine learning, in: Machine Learning: Concepts, Methodologies, Tools and Applications, IGI Global, 2012, pp. 535–578.

[23] H. Hassanzadeh, M. Keyvanpour, A machine learning based analytical framework for semantic annotation requirements, arXiv preprint arXiv:1104.4950 (2011).

[24] S. Mesbah, K. Fragkeskos, C. Lofi, A. Bozzon, G.-J. Houben, Semantic annotation of data processing pipelines in scientific publications, in: European semantic web conference, Springer, 2017, pp. 321–336.

[25] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, F. Ciravegna, Semantic annotation for knowledge management: Requirements and a survey of the state of the art, Journal of Web Semantics 4 (2006) 14–28.

[26] M. Neves, J. Ševa, An extensive review of tools for manual annotation of documents, Briefings in bioinformatics 22 (2021) 146–163.

[27] M. Tallis, Semantic word processing for content authors, in: Proceedings of the Knowledge Markup & Semantic Annotation Workshop, Florida, USA, 2003.

[28] P. Andrews, I. Zaihrayeu, J. Pane, A classification of semantic annotation systems, Semantic Web 3 (2012) 223–248.

[29] P. Ogren, Knowtator: a protégé plug-in for annotated corpus construction, in: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Demonstrations, 2006, pp. 273–275.

[30] A. Kalyanpur, J. Hendler, B. Parsia, J. Golbeck, SMORE-semantic markup, ontology, and RDF editor, Technical Report, Maryland Univ College Park Dept of Computer Science, 2006.

[31] K. Petridis, D. Anastasopoulos, C. Saathoff, N. Timmermann, Y. Kompatsiaris, S. Staab, M-ontomat-annotizer: Image annotation linking ontologies and multimedia low-level features, in: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Springer, 2006, pp. 633–640.

[32] M. Erdmann, A. Maedche, H.-P. Schnurr, S. Staab, From manual to semi-automatic semantic annotation: About ontology-based text annotation tools, in: Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content, 2000, pp. 79–85.

[33] M. Vargas-Vera, E. Motta, J. Domingue, M. Lanzoni, A. Stutt, F. Ciravegna, Mnm: Ontology driven semi-automatic and automatic support for semantic markup, in: International Conference on Knowledge Engineering and Knowledge Management, Springer, 2002, pp. 379–391.

[34] R. Schroeter, J. Hunter, D. Kosovic, Filmedcollaborative video indexing, annotation, and discussion tools over broadband networks, in: 10th International Multimedia Modelling Conference, 2004. Proceedings., IEEE, 2004, pp. 346–353.

[35] S. Jain, Understanding Semantics-Based Decision Support, CRC Press, 2020.

[36] S. Soderland, D. Fisher, J. Aseltine, W. Lehnert, Crystal: Inducing a conceptual dictionary, arXiv preprint cmp-lg/9505020 (1995).

[37] E. Riloff, et al., Automatically constructing a dictionary for information extraction tasks, in: AAAI, volume 1, Citeseer, 1993, pp. 2–1.

[38] M. E. Cali, Relational learning techniques for natural language information extraction, Report AI98276 (1998).

[39] D. Freitag, Information extraction from html: Application of a general machine learning approach, in: AAAI/IAAI, 1998, pp. 517–523.

[40] S. Soderland, Learning information extraction rules for semi-structured and free text, Machine learning 34 (1999) 233–272.

[41] I. Muslea, S. Minton, C. Knoblock, Stalker: Learning extraction rules for semistructured, web-based information sources, in: Proceedings of AAAI-98 Workshop on AI and Information Integration, AAAI Press, 1998, pp. 74–81.

[42] D. Freitag, N. Kushmerick, Boosted wrapper induction, AAAI/IAAI 583 (2000).

[43] S. Jain, S. Sharma, J. M. Natterbrede, M. Hamada, Rule-based actionable intelligence for disaster situation management, International Journal of Knowledge and Systems Science (IJKSS) 11 (2020) 17–32.

[44] B. Rosenfeld, R. Feldman, Using corpus statistics on entities to improve semi-supervised relation extraction from the web, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 600–607.

[45] S. Brody, Clustering clauses for high-level relation detection: An information-theoretic approach, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 448–455.

[46] R. Bunescu, R. Mooney, Learning to extract relations from the web using minimal supervision, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 576–583.

[47] C. Choi, I. Park, S. J. Hyun, D. Lee, D. H. Sim, Mire: A minimal rule engine for context-aware mobile devices, in: 2008 Third International Conference on Digital Information Management, IEEE, 2008, pp. 172–177.

[48] B. McBride, The resource description framework (rdf) and its vocabulary description language rdfs, in: Handbook on ontologies, Springer, 2004, pp. 51–65.

[49] G. Antoniou, F. Van Harmelen, Web ontology language: Owl, in: Handbook on ontologies, Springer, 2004, pp. 67–92.

[50] S. Jain, C. Gupta, A. Bhardwaj, Research directions under the parasol of ontology based semantic web structure, in: International Conference on Soft Computing and Pattern Recognition, Springer, 2016, pp. 644–655.

[51] N. Limbasiya, P. Agrawal, Semantic textual similarity and factorization machine model for retrieval of question-answering, in: International Conference on Advances in Computing and Data Sciences, Springer, 2019, pp. 195–206.

Table 3
Ontology Support Semantic Annotation tools

SA System	Approach	Description	Application domain	Automation
AnnotEx	Supervised Learning	Annotating based on classifying documents by means of semantic similarities	General	Manual
S-CREAM	Supervised Learning	Annotate Dynamic web pages and track the activities using hyperlink	Domain Dependent	Semi-Automatic
NavEx	Supervised learning	Extends traditional performance-based annotation	Service Oriented Environments	Automatic
Knowledge and Information Management (KIM)	Supervised Learning	Keyword-based	Inter-domain knowledgebase	Manual
Armadillo	Supervised learning	Gene annotation system, Pattern Discovery	Gene (Biomedical)	Automatic
CREAM	Rule-Based/wrappers	Framework for high structure web page	General	Automatic / Manual
GoNTogle	Unsupervised	Annotation and search facilities based on textual similarity	General	Automatic
C-PANKOW	Unsupervised	Pattern based annotation task	General	Automatic
BIMTag	Unsupervised	Semantic annotation of online BIM product resources	BIM product	Automatic
OEAKM	Unsupervised	Built ontology enabled annotation KMS that provides clustering and real-time discussion for collaborative learning	General	Semi-automatic
Melita	Supervised	Follow to two phase cycle (Turning and scheduling text) based on the training and active learning	General	Manual / Automatic
OntoMat-Annotizer	Unsupervised	Web based annotation tool that is able to create owl instance, attribute and relationship	Image, Multimedia Manual	Automatic
AeroDAML	Rule based	Scalable with diverse ontologies	Webpage	Semi-Automatically
PARMENIDES	Unsupervised	create a domain ontology using cluster	General	Automatic
MnM	Unsupervised	Learns extraction rules from training corpus	webpage	Semi-Automatic
SemTag	Rule based	Performs structural analysis	General	Automatic