


Article

Development of Knowledge Base Using Human Experience Semantic Network for Instructive Texts

Hossam A. Gabbar ^{1,*},[†] , Sk Sami Al Jabar ^{2,†} , Hassan A. Hassan ³ and Jing Ren ²

¹ Faculty of Energy Systems and Nuclear Science, Ontario Tech University, Oshawa, ON L1G 0C5, Canada

² Department of Electrical and Computer Engineering, Ontario Tech University, Oshawa, ON L1G 0C5, Canada; sksamial.jabar@ontariotechu.net (S.S.A.J.); Jing.Ren@uoit.ca (J.R.)

³ IRI, Reactor Innovation, Ontario Power Generation, Whitby, ON L1N 9E3, Canada; Hassan.Hassan@opg.com

* Correspondence: Hossam.Gabbar@uoit.ca

† These authors contributed equally to this work.

Abstract: An organized knowledge structure or knowledge base plays a vital role in retaining knowledge where data are processed and organized so that machines can understand. Instructive text (iText) consists of a set of instructions to accomplish a task or operation. Hence, iText includes a group of texts having a title or name of the task or operation and step-by-step instructions on how to accomplish the task. In the case of iText, storing only entities and their relationships with other entities does not always provide a solution for capturing knowledge from iTexts as it consists of parameters and attributes of different entities and their action based on different operations or procedures and the values differ for every individual operation or procedure for the same entity. There is a research gap in iTexts that created limitations to learn about different operations, capture human experience and dynamically update knowledge for every individual operation or instruction. This research presents a knowledge base for capturing and retaining knowledge from iTexts existing in operational documents. From each iTexts, small pieces of knowledge are extracted and represented as nodes linked to one another in the form of a knowledge network called the human experience semantic network (HESN). HESN is the crucial component of our proposed knowledge base. The knowledge base also consists of domain knowledge having different classified terms and key phrases of the specific domain.

Keywords: knowledge base; entity relationship extraction; natural language processing; human experience semantic network; knowledge structure



Citation: Gabbar, H.A.; Jabar, S.S.A.; Hassan, H.A.; Ren, J. Development of Knowledge Base Using Human Experience Semantic Network for Instructive Texts. *Appl. Sci.* **2021**, *11*, 8072. <https://doi.org/10.3390/app11178072>

Academic Editor: Jesualdo Tomás Fernández Breis

Received: 2 July 2021

Accepted: 22 August 2021

Published: 31 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Instructive texts (iTexts) are different, in terms of structure and textual pattern than standard texts. iTexts usually instruct or describe how to do something in a step-by-step process. For example, how does one fix a turbine? The answer to this question has a few procedures to follow, which will help accomplish the main goal or operation. iTexts usually consist of a title, which could be the name of the process or operation, and a set of instructions or procedures that help to accomplish the operation in a step-by-step process. Figure 1 shows the differences between regular or standard text and iText in terms of structure and textual pattern.

Employees in large industries, such as the nuclear industry, store information like procedures, precautions, experiments, risk factors, etc., in handwritten or pdf documents, which are prominent in quantity. These are called operational documents. They follow these documents during operation in order to accomplish each task efficiently. There is a continuous movement of experienced personnel to different departments, or they go for retirements and hence a tremendous amount of expertise is lost. The loss of expertise costs the industry a huge amount of money as they have to invest in training less experienced personnel, leading to indirect losses in delayed or wrong activities. A less experienced

employee cannot operate complex tasks due to having less knowledge and training about the documents and their operation. The training period could take months to cover information about the different operations. The more extended the training period, the more expensive it is for the industry. At many times, it is troublesome to retrieve any specific information during operation or other practices. It is helpful if the desired information is quickly retrieved when employees are in the middle of an industrial activity or in a lab, making them work faster. Moreover, much time is wasted while searching for specific information from one out of innumerable documents during a complex operation to accomplish its objective. In case of any inaccurate information retrieval, there is a high risk of operational failure, which is again costly to recover for the industry. If information and human experience from these large number of documents could be extracted, structured and retained in a knowledge base from where desired information could be easily retrieved at any time, then the operational time could be saved and utilized in a much better way. Furthermore, this could also reduce the expenses for the training and learning purpose. The learning process could also be faster. The less experienced employee will also be able to perform the complex operation with the help of the knowledge base, which was impossible for them previously. However, the management of this knowledge base could be critical with the increase in information. Without proper structuring of knowledge, information retrieval will be an expensive approach.

Regular Text

Canada is a country situated in North America. It has 10 provinces and 3 territories. It is the second largest country in the world. The capital of Canada is Ottawa and Toronto is the largest city.

Bangladesh is a country in South Asia. Dhaka is the capital and largest city. Its population is 163 million. Bangladesh is one of the emerging and growth-leading economies of the world, and is also one of the Next Eleven countries, having Asia's fastest real GDP growth rate.

Instructive Text

Handling product "A" to prevent damage:

1. Wear gloves.
2. Use mask when handling the product.
3. Leave protective caps and covers on the product until installation.
4. Always keep "A" under water before leaving the lab.

Handling product "B" to prevent damage:

1. Wear gloves.
2. Use mask when handling the product.
3. Keep product "B" always under water.
4. Maintain the temperature of water 37 degrees.

Figure 1. Difference between regular text and instructive text (iText).

Hence, this knowledge can be structured in the form of a network, being able to retain the human expertise from these documents in an organized way by developing relationship among the entities, their action, attributes and different values and parameters from each of the iTexts and procedures, which could have information about human role, tool, equipment, location, document, operation, procedure, etc., associated with that particular operation. Current research approaches in developing a knowledge base and retaining the relationship among entities and structuring knowledge from standard texts does not fully apply for iTexts as relationships, attributes, and properties of entities in iTexts differ in case of different operation. This paper presents a knowledge base consisting of HESN, which structures knowledge by capturing the human experience from iTexts using grammar, semantic meaning, and domain knowledge consisting of classes and properties of information related to that specific domain. The knowledge base retains the properties, relationships and values of different entities, action terms or verbs, attributes, and attribute values found in an iText for different operations. It extracts the real expertise from iTexts and dynamically updates the HESN existing in the knowledge base. The contribution of this work can be summarized as follows:

1. The development of an adaptive, dynamic and deterministic knowledge structure with qualitative and quantitative attributes, called the human experience semantic network (HESN), is used to capture and structure knowledge from iTexts in the form of nodes and edges;

2. The development of a knowledge base, consisting of HESN and domain knowledge, for retaining properties, values, and relationships of different terms or key phrases, found in iTexts. These terms or key phrases could be an entity, action term or verb, attribute, or attribute value. The knowledge is structured for different entities, action terms, attribute, or attribute values based on different operations.

The rest of this paper is organized in the following way. In Section 2, a literature review is done related to the work. In Section 3, the methodology and HESN is explained. Section 4 demonstrates the advantage and ability of this approach. A conclusion is drawn in Section 5.

2. Related Work

There are many ways to capture knowledge from text and develop a knowledge base, knowledge network, or knowledge graph to represent the acquired knowledge. This section discusses research work related to the knowledge base, information extraction approaches, and entity-relationship establishment approaches to provide information about recent work on how knowledge is acquired and represented from different kinds of texts based on different domains. One major part of our proposed knowledge base consists of information about the relationships of different terms and key phrases in iText based on different operations. All these relations are represented in the knowledge network called HESN. Hence, the literature review includes research works on both knowledge base techniques and entity-relationship extraction techniques to represent knowledge.

2.1. Knowledge-Based and Ontology-Based Approaches

The system, which is based on a knowledge base, consists of information and data structured in an organized way. Question answering over a developed knowledge base based on the domain knowledge helps retrieve the information as demanded through a query. It is necessary to develop a dynamic knowledge base that can capture knowledge and dynamically update the knowledge base in such a case. Knowledge-based approaches are also used for learning and extracting knowledge from texts based on a specific domain. In [1], a method has been proposed for the development of a knowledge base based on the knowledge and behavior of operators in terms of a severe accident in nuclear power plant. The knowledge base was developed using 281 scientific publications, which were summarized and then the knowledge was extracted from them. The publications were related to the terms “severe accident” and “nuclear”. The knowledge base that constitutes the knowledge graph consists of nodes and edges representing the causal relationships, entities, states, and affiliations. A knowledge graph was generated from each publication summary, and all these knowledge graphs were merged to constitute one main knowledge graph. From this methodology, it is observed that texts are skipped because it is considered repetitive or irrelevant to the topic of interest. This skipping of text is unacceptable in dealing with iTexts as each and every information in each of the instructions is required to be extracted and structured correctly in the knowledge base. Moreover, procedural knowledge structuring, information tracking, and sequencing are a significant part of structuring knowledge from iTexts which cannot be solved only with entity relationships. In [2], an agricultural knowledge base or framework has been proposed which helps to identify pests and diseases that affect a crop. An automatic ontology population tool has been developed. It helps extract relevant data from unstructured documents with the help of natural language processing techniques and update ontology. Their approach included representing information related to symptoms of plant diseases based on plant parts and damages. The proposed methodology was built into a system that could recognize crop pests with the help of the knowledge base. A system called “Smart Farming” was proposed in [3] for precision farming management, where a knowledge base was developed. Information was organized in the knowledge base in the form of a semantic network consisting of concepts and relations. The knowledge represented in the knowledge base was about crop production, production resources, agricultural machinery, equipment,

and other resources. Ontological principles were adopted to design the domain model based on concepts, attributes, and interrelations. In [4], a knowledge-based strategy has been proposed for data management and mining in machining and to support decision making. The knowledge base consisted of manufacturing knowledge and a multi-level model that helped acquire knowledge and decision-making from the information stored in the knowledge base. Operation optimization knowledge base system (OOKBS) was designed in [5] for the operation optimization of a polyethylene process. Knowledge was represented using an ontology. Knowledge of polyethylene process, equipment operation, and operation optimization were integrated into the knowledge base. A neural network model was developed to identify the relationship between operating conditions and molecular weight distribution (MWD) parameters.

In [6], an ontology-based approach was proposed, which classifies security requirements automatically. The security requirements were described with the help of 35 defined linguistic rules and 140 defined security keywords. The security requirements ontology was defined using description logic (DL). All these are used to train classifiers of security requirements using machine learning algorithms like naïve Bayes (NB), decision tree (DT), and logistic regression (LR). In [7], an ontology, named concrete bridge rehabilitation project management ontology (CBRPMO), was presented, which was developed using domain knowledge of bridge rehabilitation and following standard procedures. Semantic reasoning rules were constructed to support dynamic information integration and management functions. The developed ontology aims to investigate the information in bridge rehabilitation projects and efficiently support constraint management. The ontology was developed using web ontology language (OWL). A knowledge-based model for additive manufacturing (AM) has been proposed in [8] using ontology, where data are organized with the help of the ontology structure. A form is filled up with data and based on that, data validation and reasoning are done with the help of associated rules that determine the appropriate machine name or model that can do the manufacturing. Moreover, the paper [9] proposes a knowledge-based approach that covers different ontology learning methods from the text.

2.2. Entity-Relation Extraction

The main objective of the entity recognition and relation extraction task is to determine the relational structure of the mentioned entities from unstructured texts. The task has two subtasks, namely (i) named entity recognition (NER) [10] and (ii) relation extraction (RE) [11]. These tasks help to connect each entity with other ones and are very useful for developing a semantic network with nodes and edges. In this research work, our knowledge base help to identify the entity and key phrases from iTexts and represent the relationship among them based on different operation in the form of HESN. In [12], an approach has been proposed for constructing knowledge graphs with the help of a task named relation extraction and linking. Their approach is dependent on information extraction (IE) tasks for obtaining the named entities and relations. Finally, these are linked using data and standards of the semantic web. Initially, the input text is transformed into resource description framework (RDF) triples using the combination of natural language processing and information extraction operation. The information extraction operation includes tasks like document acquisition and preprocessing the input text, extracting named entities and their association with the grammatical unit, semantic relation extraction using the OpenIE approach and associating it with semantic information provided by an approach termed as semantic role labeling (SRL) that helps to identify the order and selection of elements which are to be finally represented through RDF triples. In total, 605 IT news webpages were downloaded and used for the evaluation of their research methodology. It had about 12,015 sentences which were processed to construct the RDF statements. RDF triples or statements are dependent upon the subject, object, and predicate of a sentence. Many sentences were ignored that had relations containing no named entities in subject and object. As a result, RDF statements for such sentences were not created

and, thus, ignored. In [13], a relation extraction method for construction of COVID-19 information knowledge graph based on deep learning was proposed. It is another open information extraction (OpenIE) system based on unsupervised learning without any pre-defined dataset, although a COVID-19 entity dictionary was created and used for scraping related information. The proposed method extracts knowledge from documents consisting of information related to COVID-19 and constructed a knowledge base that consisted of connecting words between COVID-19 entities, which was captured from COVID-19 sentences. The proposed model could identify a relation between COVID-19-related entities using (BERT), and it does not need any pre-built training dataset. In [14], researchers presented a neural model to extract entities and their relation from texts. The basic layers of their proposed model consisted of embedding layer, bidirectional sequential long short term memory (BiLSTM) layer, conditional random fields (CRF) layer, and sigmoid layer. The conditional random fields (CRF) layer was used to recognize entities and the sigmoid layer for entity relation extraction. The task was modeled as a multi-head selection problem where an entity may have multiple relations in a text. The model does not rely on hand-crafted or external natural language processing tools, such as parts-of-speech (POS) tagger, dependency parsers, etc. Extracting semantic relation from text has been performed by a group of researchers in [15]. Two models named Rel-TNG and Type-TNG were proposed that used topic n-Grams (TNG). These two models were able to show similar performance measure for Rel-LDA and Type-LDA, but the models outperformed Rel-LDA and Type-LDA when there was prior knowledge available. GENIA and EPI datasets were used for this experiment which are biomedical texts. Two types of relationships were annotated: PROTEIN-COMPONENT and SUBUNIT-COMPLEX. One of their advantages was that these annotations were already done and were provided with the dataset. Ref. [16] shows another knowledge graph construction mechanism that involves entity-relationship establishment from triples consisting of entities and their relationships and also fulfilling relationship gaps between entities from texts containing those relationships. The approach uses texts to fulfill relationship gaps found between entities. The knowledge graph does not capture the relationship between texts as their approach is not aimed to capture that. They aimed to find and identify triples (h, r, t) , where h and t are entities, and r is the relationship, and extract relationship from there and use texts to find if it is missing between any h and t . An investigation is done in [17] which narrated the influence of semantic link networks on the performance of the question answering system. It is accomplished by enhancing the ability of the system in answering different types of questions and supporting different patterns of answering questions with the help of the semantic link network. The accuracy of an answer against a question depends widely on the answer range and the number of semantic links on the answer range. By answer range, it has been meant to have more texts having potential answers. The research work clarifies that the greater the number of semantic links there is, the accuracy and formation of the answer will be better against a question. The semantic link network is formed from semantic objects and their semantic links that connect two semantic objects. These objects consist of a form of a string with their synonyms. Semantic link network, produced by the researchers, connects different terms from a range of text and establishes relationships, which is almost similar to entity relationship establishment.

2.3. Limitations in Case of iText

Different methods and approaches of information extraction and knowledge structuring discussed are suitable for learning from regular or standard texts or paragraphs consisting of information about different entities and their relationships. Relationships among entities were established in the form of RDF triples, semantic networks or knowledge graph. Knowledge extraction is performed from texts and developed knowledge-base. Ontology is developed for better success of the knowledge extraction process. Semantic meaning is extracted with the help of ontology. Identification of different problems is performed based on established knowledge-base or ontology. Such approaches have

limitations in capturing knowledge from iTexts that consists of a set of instructions related to how to conduct an operation or activity. The structure of the sentence in iText is a bit different from regular paragraphs. Firstly, in iTexts, there could be an entity having different values or relationships based on different operations. Hence, it is essential to keep track and structure knowledge of the values and relationships of an entity based on different operations. Secondly, there could be relationships between two entities and between an entity and other terms like “move”, “shift”, “high”, “low”, or any number. Traditional triplets extraction or RDF triples are extracted from sentence structure consisting of subject, object, and predicate. This method is not perfect as it sometimes consists of error or contradictory information [18]. Moreover, having more than 1 triple in a sentence is again required to be handled which is also an expensive process [19]. Predicting triple in a sentence is also another task that needs effort. It is also not always possible to get information in the form of triplets. In Figure 1, the iText “Wear gloves” consists of two words only. Here, a particular user could be considered as subject. However, in case of iText, this is not always applicable. Considering “user” as subject and relating this entity with other two terms will create confusion in the case where this “user” is already defined as any particular human role in the operation title. Therefore, if the operation title is “Must perform tasks for Lab Operators”, then wearing gloves is instructed for lab operators instead of “user”. Duplet based relation extraction helps to create relation between the action “wear” and entity “gloves”. This piece of information also consists of tags from operation title that helps to know that wearing gloves is applicable for lab operators. This tracking of information is explained in more detail in methodology part of this paper. Moreover, ontology or domain knowledge in case of RDF needs to be enriched. However, in our case, the domain knowledge is developed in a simple way. It consists of different class names and words or phrases defined under each class name, which is good enough for generating duplet relations. Further explanation about domain knowledge is done in the methodology part of this paper. Another example can be drawn when numbers are considered. In case of duplets, our approach identifies number and can make relation with another term or phrase directly, whereas, it is a complicated task when triplets are considered. For example, “Pump must have pressure 4 Pa”. From here, we get the duplets (pump, pressure), (4, pressure). Here, the value 4 is directly assigned with pressure which will be an helpful information for complex reasoning when this operational instruction is considered. This is not possible with triplets directly. For this reason, our research deals with duplets. The domain knowledge also needs to be well defined. This will help to identify different terms found in iTexts and structure knowledge accordingly. This paper proposes a knowledge base that captures knowledge from iTexts, represents the knowledge using HESN as part of the knowledge base, and dynamically updates the knowledge base.

3. Research Methodology

The development of the knowledge base for iTexts is a step-by-step process. The two major parts of the knowledge base are domain knowledge and HESN itself. They help accomplish tasks, such as identifying different terms and key phrases, establishing relationships among them, structuring knowledge of different terms, and key phrases found in iTexts based on different operations under which each of the instructions is provided and finally update HESN and the knowledge base. For simplicity, all entities and named entities are termed entities in this paper. In this research, a relationship is established among four types of terms or key phrases—entities, action terms or verb terms, attribute terms, and attribute values. Domain knowledge consists of information related to these terms or key phrases. They are represented using class and property and help to detect and identify terms and phrases in iTexts. Each time new instructions are learned, and the HESN is updated. Updating HESN or domain knowledge also means updating the knowledge base since HESN and domain knowledge constitute the knowledge base.

3.1. iTexts Extraction and Preprocessing

This research is done based on the test documents communicated with the maintenance section within Ontario power generation (OPG), responsible for approximately half of the electricity generation in the Province of Ontario, Canada. The test documents had different contents related to purpose, pre-requisites, instructions, post-requisite, definitions, summary of changes, validation, and verification, and similar information about different processes, operations, inspections, equipment, etc. The texts related to operational procedures and instructions, only those texts were extracted with the help of an algorithm that was developed to follow a standard procedure and group the texts combining related sentences. In this way, the entire document is divided into small chunks where each chunk consists of the title of the operation or procedure (Parent-iText or PT) and a set of instructions underneath (Child-iText or CT). These groups or chunks of texts were further processed to capture knowledge and retain it in the knowledge base with the help of HESN and domain knowledge. However, this paper focused mainly on extracting knowledge from iTexts and developing the knowledge base and HESN rather than text extraction from documents.

3.2. Domain Knowledge Development

The domain knowledge is an essential part of the knowledge base. It is used to identify the entities, action terms, and attribute terms and values from the iTexts. The entities are the nouns or names of different equipment, document, tool, human role, etc. The attribute terms could be terms like height, pressure, weight, etc. Attribute values refer to any number or value, and it could be status, such as complete, in progress, condition of anything, such as poor, high, dry, and so on. The term “action” refers to the verbs, such as measure, work, move, check, etc. Each entity, action, attribute, and attribute value, except for the numbers, are pre-defined and classified as part of the knowledge base. There could be many classes, and under each class, there exists multiple entities, action, attribute, or attribute values. Each of these classes has properties. For example, “humanRole” is the name of a class. Under this class, there are entities such as “engineer”, “personnel”, “manager”, etc. The domain could be related to a nuclear power plant, chemical industry, or any other category. Based on this domain, different terms and phrases are defined and classified with the help of domain experts. The more enriched domain knowledge is, the more term identification is possible. As a result, more relationship establishment is possible later on among different terms found in the iText. The properties of each class help to know about a term or phrase’s association with other terms and from which iText and operation, the term was identified. This information about different terms or phrases is updated each time when new iTexts are read from the document.

3.3. Human Experience Semantic Network (HESN)

HESN is the key component of our proposed knowledge base. Different terms and key phrases are identified from iTexts with the help of domain knowledge. There could be different operations or procedures in a document. Under each operation, there could be multiple instructions or procedure that talks about how to accomplish that particular operation. There could be the same term or key phrase in different operations. HESN represents the knowledge network that shows the association or relation of a term or key phrase with other terms or key phrases based on different operations. Each of these terms or key phrases could be an entity, action, attribute or attribute value. The network is represented in the form of nodes and edges that constitute a tree or undirected graph. Figure 2 represent a small glimpse of HESN where nodes and edges are connected. Figure 3 represents detailed information about each node. As the network retains the relations, semantics, and information about different terms and key phrases, and captures the knowledge and experience from iTexts existing in operational documents, thus it is called the human experience semantic network (HESN). Creating relations among different terms or key phrases based on the operation is performed with the help of tags. The methodology

of creating relations among different terms and phrases and the use of tags is explained in the latter part of this paper.

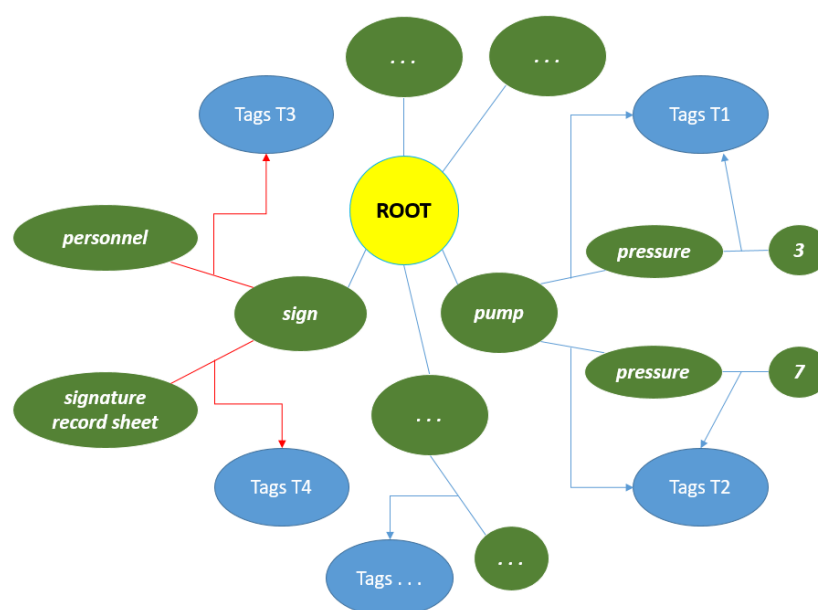


Figure 2. Human experience semantic network (HESN).

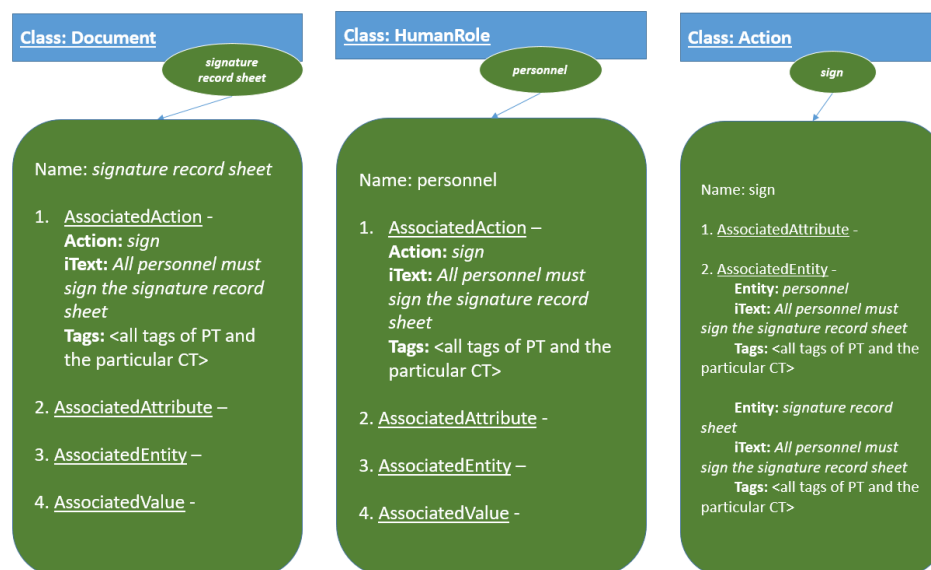


Figure 3. Three different entity and their classes and property found in domain knowledge. Values are updated when new iText is read.

3.4. Entity, Action, Attribute and Value Recognition and Linking

Domain knowledge is used to deal with recognizing terms and key phrases, which could be an entity, action term, attribute term, or some value. If named entities, action, or attribute, that consists of more than one word are identified, each word of that named entity, action, or attribute is concatenated to make it a single word. For example, water pump = waterpump. This helps in making the relationship among the words or key phrases easier later on. An attribute could be terms like pressure, height, condition, etc., which are the properties of an entity. Its value could be high, low, poor, etc. It could also be a numeric value. The domain knowledge consists of all these terms, except for the numeric values. Once the identification and concatenation are made, the next task is to establish relationships among the words or phrases. At first, the stop words are re-

moved from the sentence except for a few, which are “on,” “in,” “this,” “have,” “has,” and “should.” There could be six types of relationships—(i) entity-action (E-Ac), (ii) entity-entity (E-E), (iii) entity-attribute (E-Att), (iv) entity-value (E-V), (v) action-attribute (Ac-Att), and (vi) attribute-value (Att-V). The relationship is always created among two terms or key phrases. A grammar pattern-based linguistic matching is done with the help of a library named spaCy [20]. This helps to identify the direct dependency of a word over another word in a sentence in the form of a duplet. Each of these duplets is further processed and reorganized.

Figure 4 shows the algorithm using which the tags are created and duplets are generated from each iText. Tags are the nouns and verbs found in an iText. A set of tags are used against each duplet. It helps to identify from which particular iText, the duplet was generated. Furthermore, this information helps to distinguish the relationships between different terms and key phrases based on different operations. The use of tags is explained further in the latter part of this paper.

- Step 1: start
- Step 2: read $T[i]$ from OP
- Step 3: tokenize $T[i]$ using NLP
- Step 4: extract N and V from $T[i]$
- Step 5: $TAGS = [N, V]$
- Step 6: if $T[i] = CT$,
 $ALLTAGS = TAGS + PTAGS$, Go to Step 8
- Step 7: else if $T[i] = PT$, $PTAGS = TAGS$, Go to Step 16
- Step 8: $sp = \text{spacy.load('en_core_web_lg')}$
- Step 9: $T[i] = \text{removeStopWords}(T[i])$
- Step 10: $doc = sp(T[i])$
- Step 11: $DD = \text{getAllDD}(doc)$
- Step 12: for d in DD –
 if ($d[0]$ or $d[1]$ does not contain DK or numeric value), remove d from DD
- Step 13: for d in DD –
 if ($d[0]$ does not contain DK or numeric value),
 for w in DD –
 if ($d[0]$ in w and $d[1]$ not in w)
 if $d[0] = w[0]$, $d[0] = w[1]$
 else $d[0] = w[0]$
 break
 else if ($d[1]$ does not contain DK or numeric value),
 for w in DD –
 if ($d[1]$ in w and $d[0]$ not in w)
 if $d[1] = w[0]$, $d[1] = w[1]$
 else $d[1] = w[0]$
 break
- Step 14: for d in DD –
 if ($d[0]$ and $d[1]$ does contain DK or numeric value), remove d from DD
- Step 15: $\text{updateHESN}(DD, ALLTAGS)$
- Step 16: $i = i + 1$
- Step 17: if $T[i]$ exist, go to step 2
- Step 18: else stop

Figure 4. Algorithm of creating tags and duplet formation.

From the algorithm, OP in Step 2 refers to a set of instructions having a title or operation name (PT) and one or more instructions (CT). $T[i]$ represents each iText which could be a PT or CT. ‘N’ and ‘V’ in the algorithm means all nouns and verbs extracted from that particular iText. ‘TAGS’ in Step 5 denote all the Nouns and Verbs of $T[i]$, whereas “ALLTAGS” in Step 6 denote tags of that particular iText and the ‘PTAGS’. ‘PTAGS’ are the tags extracted from PT. All necessary components of the spaCy library is loaded and assigned to ‘sp’ in Step 8. It can now be used to perform tasks like finding word dependencies from within a sentence. In Step 9, the stop words are removed from the

iText except for a few, which are ‘on’, ‘in’, ‘this’, ‘have’, ‘has’, and ‘should’. In Step 10, the iText is processed using ‘sp’ to get valuable insight, such as direct word dependencies, parts of speech tag for each word, etc. In Step 11, the function “getAlIDD” returns word dependency for each word in the sentence in the form of duplets. Each element in the duplet is represented as d[0] and d[1], as shown in Step 12. “DK” consists of all terms found in domain knowledge. The final “DD” found in Step 14, after ending the loop, consists of the sorted duplets. Concatenation of the duplets creates a small network for that particular iText, as shown in Figure 5. This network is the building block of HESN. Figure 5 visually represents the methodology of how a small HESN network is generated from an iText. Step 15 is described in section “Update HESN” of this paper.

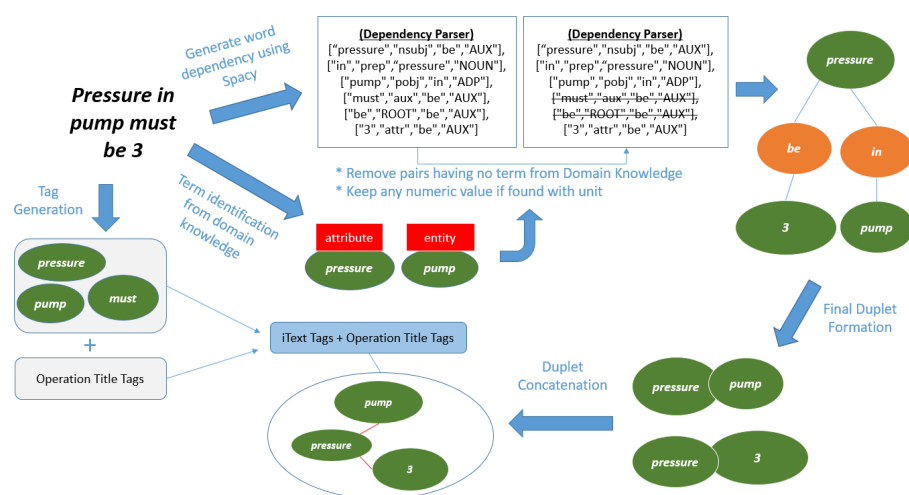


Figure 5. Generation of duplets and formation of small network from iText.

3.5. Tag Generation and Relation Tracking

When it comes to iTexts, it is essential to track the information about different terms and phrases provided in different sets of instructions or operations. If we again consider Figure 5, we get the entity here as “pump”, and its attribute is “pressure”. The value is mentioned as 3. Let us consider this value for ‘pump’ for operation OP1. There could be another operation OP2 where the entity and attribute are the same but the value is 7. In this case, two different values are obtained having the same attribute of the entity but for different operations, OP1 and OP2. In order to keep track of this knowledge, tag plays an important role. Figure 6 shows how relations of the same entity are structured for two different operations. Tags are termed in this research as the nouns and verbs extracted from text, having word’s character length greater than 2 for verbs and any character length for nouns. For every network that is generated from each instruction, tags are added against them. These tags contain the nouns and verbs extracted from that particular instruction and the title of the operation under which the instruction is situated. Considering the same example from Figure 5, if T1 is considered as the set of tags for those associations found in the small network, formed from that particular iText, then T1 consists of the nouns and verbs of that iText (CT), along with tags generated from the title of its operation (PT). This takes place for every instruction under the same operation. This helps to keep track of which information is coming from which operation. Parts-of-speech (POS) tagging is one of the popular techniques of natural language processing. It has been used in this research for generating the tags from each iText. The process is shown in Figure 7.

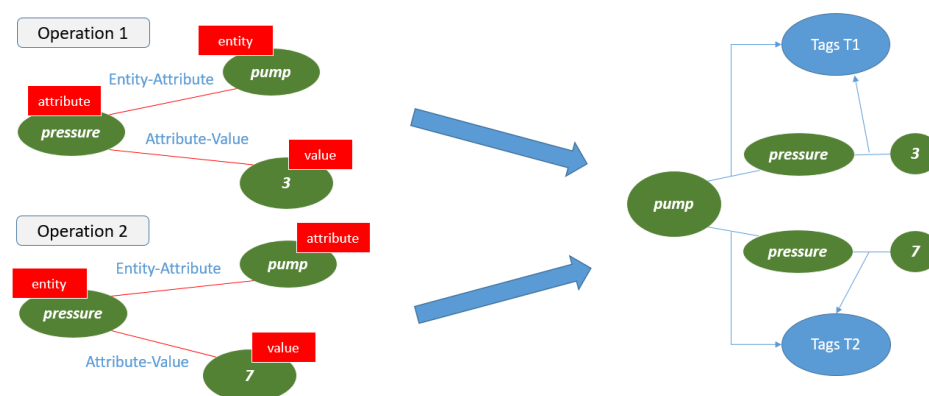


Figure 6. Updating value of same entity from two different iText for two different operation which shows how HESN is updated.

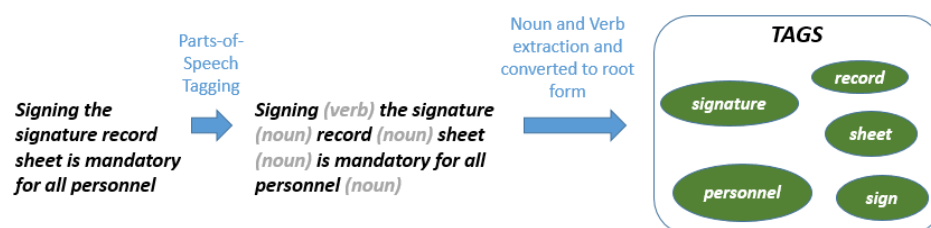


Figure 7. Extracting nouns and verbs from text and converted to root form as tags.

3.6. Update HESN

From Figure 5, it is observed how a small network is generated from each iText consisting of the relationship among terms of entity, action, attribute, and its value and how the respective tags are generated from that particular instruction (CT) and title of the operation (PT). HESN consists of nodes and edges. Figure 3 represent detailed information about a node. Whenever a new relation is created between two terms or key phrases, the property of both of the terms are updated. For example, in Figure 3, the term “personnel” and “sign” are related and was found from an iText. The term “personnel” is an entity whereas the term “sign” is an action. The property “AssociatedAction” of “personnel” is updated with the term “sign”, which indicates that the entity “personnel” is related with an action called “sign”. It also includes the iText, where these terms were identified. Moreover, it also stores the information extracted as tags from title of the operation. This gives an idea about the operation under which the iText was found. If a new term, such as “move”, is found related with “personnel” from a new iText, then the new term “move” is added to the property “AssociatedAction” of the term “personnel” in the similar way. Every new association creates a small network. All these small networks together form a more extensive network, which is the HESN. This is how each node of HESN and the network itself is updated. As the information of HESN is stored in the knowledge base, the knowledge base is also updated.

4. Advantage of the Proposed Knowledge Base

The structuring of knowledge from iTexts and capturing the human experience from it is not only about extracting or establishing relationships among the entities found in each iText, but more about linking that information and relation with different operations, which is done using tags, as shown previously in the methodology. The six types of relation which are established from iTexts are—(i) entity-action (E-Ac), (ii) entity-entity (E-E), (iii) entity-attribute (E-Att), (iv) entity-value (E-V), (v) action-attribute (Ac-Att), and (vi) attribute-value (Att-V).

4.1. Query Evaluation

The knowledge base proposed in this paper is advantageous when learning from operational and procedural documents that consist of iTexts. The information observed in operational documents, consisting of iTexts, needs to be retrieved based on different operations. Status, condition, involvement of human role, measurement, activity, etc., varies for different operations, although the terms are the same. Hence, when a query is asked based on an operation, HESN can provide information according to that particular operation. This makes HESN unique and efficient for iTexts. From Figure 6, two queries could be considered:

1. What should be the pressure of pump for Operation 1?
2. What should be the pressure of pump for Operation 2?

Here, “Operation 1” and “Operation 2” are the title (PT) of two separate operations. If the tags of CT of “Operation 1” is T1 and tags of CT of “Operation 2” is T2, respectively, then it is possible to retrieve the network consisting of the relation among “pressure”, “pump”, and “3” based on T1 and the network consisting of the relation among “pressure”, “pump” and “7” based on T2. In this way, both the questions can be answered using HESN. Moreover, domain knowledge consists of information about the classes of each of the terms. This helps identify entities, actions, attributes and attribute values, and complex reasoning through HESN.

4.2. Relation Extraction

The accuracy of relation extraction is measured based on the procedural and operational test documents provided by OPG. In total, 25 different types of sentences or iTexts were selected, and 102 relations were extracted. Each relation is made between 2 keywords or phrases. A total of 16 relations were ignored as they do not fall into previously mentioned six types of relations, and 79 relations were correctly extracted. Figure 8 is a table that shows what duplets are generated from each iText. Figure 8 is shown to provide an example of how duplets are generated and finalized from each iText. Each of these duplets contains a relation, and the terms are already classified in the domain knowledge. In the “relation” column, “TRUE” means that a particular duplet follows one of the six types of relations that were previously mentioned, and “IGNORED” means it does not follow. “FALSE” means the relation of the duplet is wrong. ‘E’, ‘Ac’, ‘Att’, and ‘V’, that is observed in “duplets” column in Figure 8, stands for “Entity”, “Action”, “Attribute”, and “Value”, respectively. Combination of each of these duplets for a particular iText forms a network that is the building block of HESN. Each of these networks is tracked with the help of tags. Finally, the information for each entity, action, attribute, and value is updated, which updates HESN and the knowledge base as a whole.

iText	duplets	relation	wrong	correct/total
personnel must wear suit in radiation lab	personnel (E), wear (Ac)	TRUE	0	3/3
	suit (E), wear (Ac)	TRUE		
	radiation lab (E), wear (Ac)	TRUE		
prerequisites have been completed	Prerequisites (E), completed (Ac)	TRUE	0	1/1
record channel number, date, start time and repositioning required in appendix B datasheet 1	record (Ac), channel number (Att)	TRUE	0	3/3
	date(Att), channel number (Att)	IGNORED		
	start time (Att), repositioning (Att)	IGNORED		
	repositioning (Att), channel number (Att)	IGNORED		
	appendix b (E), datasheet 1 (E)	TRUE		
all personnel working on this procedure have signed the signature record sheet attached in Section 2.5 of this document	datasheet 1 (E), repositioning (Att)	TRUE	0	3/3
	personnel (E), signed (Ac)	TRUE		
	signature record sheet (E), signed (Ac)	TRUE		
	section 2.5 (E), signature record sheet (E)	TRUE		
FLM independently verify the FME requirements are established	FLM (E), verify (Ac)	TRUE	0	4/4
	FME (E), requirements (E)	TRUE		
	requirements (E), verify (Ac)	TRUE		
	established (V), requirements (E)	TRUE		
tool tethering is mandatory. Also ensure that the catch tray is in place for all repositioning activities	Tool (E), tethering (Ac)	TRUE	0	2/2
	mandatory (V), tethering (Ac)	IGNORED		
	catch tray (E), ensure (Ac)	TRUE		
	repositioning (Att), ensure (Ac)	IGNORED		
execution cart is setup and available on both platforms to view specific items as required via custom mounted cameras	execution cart (E), setup (Ac)	TRUE	1	3/4
	setup (Ac), view (Ac)	IGNORED		
	platforms (E), setup (Ac)	FALSE		
	items (E), view (Ac)	TRUE		
ensure equipment, listed in section test equipment tools and consumables, are prepared and ready for use	custom mounted cameras (E), items (E)	TRUE	0	4/4
	ensure (Ac), equipment (E)	TRUE		
	equipment (E), prepared (Ac)	TRUE		
	test equipment (E), consumables (E)	TRUE		
record tooling calibration data in appendix c	consumables (E), equipment (E)	TRUE	0	2/2
	ready (V), prepared (Ac)	IGNORED		
	record (Ac), tooling calibration (E)	TRUE		
	appendix c (E), tooling calibration (E)	TRUE		
ensure Communicationlinks are setup and tested between Vault and IRI trailer	communication links (E), setup (Ac)	TRUE	0	3/3
	setup (Ac), tested (Ac)	IGNORED		
	tested (Ac), ensure (Ac)	IGNORED		
	vault (E), iri trailer (E)	TRUE		
perform function test on both measuring tools, check for binding and smooth operation	iri trailer (E), tested (Ac)	TRUE	0	5/5
	perform (Ac), check (Ac)	IGNORED		
	function test (E), perform (Ac)	TRUE		
	measuring tools (E), perform (Ac)	TRUE		
ensure REP reviewed and initialled	binding (V), operation (E)	TRUE	0	1/1
	smooth (V), operation (E)	TRUE		
	operation (E), check (Ac)	TRUE		
	REP (E), reviewed (Ac)	TRUE		
FLM verify all steps completed to this point	reviewed (Ac), ensure (Ac)	IGNORED	0	3/3
	initialled (Ac), reviewed (Ac)	IGNORED		
	FLM (E), verify (Ac)	TRUE		
	steps (E), verify (Ac)	TRUE		
pump must have pressure 4 Pa	completed (Ac), steps €	TRUE	0	3/3
	pump (E), pressure (Att)	TRUE		
	pressure (Att), pump (E)	TRUE		
	4 (V), pressure (Att)	TRUE		
return the crane to the parked position when not in use	return (Ac), crane (E)	TRUE	0	3/3
	crane (E), parked (V)	TRUE		
	position(Att), parked (V)	TRUE		
	FME (E), field conditions (E)	TRUE		
FME field conditions have been reviewed and discussed with FLM	field conditions (E), reviewed (Ac)	TRUE	0	4/4
	discussed (Ac), FLM (E)	TRUE		
	FLM (E), reviewed (Ac)	TRUE		
	ensure (Ac), lubricated (Ac)	IGNORED		
ensure positioning assembly hardware, quick locknut, p/a stud threads (tube sheet end) and saddle clamp threads have been lubricated with crc penetrating oil	positioning assembly hardware (E), ensure (Ac)	TRUE	2	3/5
	quick locknut (E), positioning assembly hardware (E)	FALSE		
	quick locknut (E), positioning assembly hardware (E)	FALSE		
	saddle clamp threads (E), lubricated (Ac)	TRUE		
ensure personnel, required to access platform, has appropriate current fall arrest qualification	crc penetrating oil (E), lubricated (Ac)	TRUE	0	5/5
	personnel (E), ensure (Ac)	TRUE		
	ensure (Ac), personnel (E)	TRUE		
	access (Ac), platform (E)	TRUE		
ensure stud measurement tool has a valid calibration date	platform (E), personnel (E)	TRUE	0	3/3
	fall arrest (E), ensure (Ac)	TRUE		
	stud measurement (E), ensure (Ac)	TRUE		
	valid (V), calibration date (Att)	TRUE		
remove insulation ring on the target channel	calibration date (Att), stud measurement (E)	TRUE	0	2/2
	insulation ring (E), remove (Ac)	TRUE		
	target channel (E), insulation ring €	TRUE		
	shift engineer (E), verify (Ac)	TRUE		
shift engineer verify record of target site and confirm tool is secured in position	verify (Ac), site (E)	TRUE	2	5/7
	record (Ac), site (E)	FALSE		
	site (E), confirm (Ac)	FALSE		
	tool (E), confirm (Ac)	TRUE		
on the east target site install reconfiguration panlut and unlock target site	secured (V), tool (E)	TRUE	0	4/4
	position (Att), secured (V)	TRUE		
	site (E), install (Ac)	TRUE		
	east (E), site (E)	TRUE		
ensure the torquewrench is set to 180 ft lb	reconfiguration panlut (E), install (Ac)	TRUE	0	3/3
	unlock (Ac), install (Ac)	IGNORED		
	site (E), unlock (Ac)	TRUE		
	torquewrench (E), set (Att)	TRUE		
examine digital indicator on the e face positioning tool to determine any movement of the target channel	set (Att), ensure (Ac)	TRUE	1	3/4
	180 (V), set (Att)	TRUE		
	examine (Ac), determine (Ac)	IGNORED		
	digital indicator (E), examine (Ac)	TRUE		
checking the west digital indicator, shift the channel to the east until the required distance has been reached	e face positioning tool (E), examine (Ac)	FALSE	1	4/5
	movement (Ac), channel (E)	TRUE		
	channel (E), determine (Ac)	TRUE		
	checking, reached	IGNORED		
	west digital indicator (E), checking (Ac)	TRUE	1	4/5
	shift (Ac), channel (E)	TRUE		
	channel (E), east (E)	TRUE		
	east (E), distance (E)	FALSE		
	distance (E), reached (V)	TRUE		

Figure 8. Relations extracted from different types of sentences or iTexts.

5. Conclusions and Future Work

Knowledge extraction from iText is not similar to that from regular text or paragraph. For iTexts, it is imperative to structure information and relationships of a term or key phrase with other terms based on different operations. This research work proposes a knowledge base that helps to capture the knowledge or human experience from iTexts and dynamically update the knowledge structure. HESN and domain knowledge are the two parts of our proposed knowledge base. Domain knowledge is used to identify different terms from the iTexts. HESN is used to represent the knowledge from iText. This knowledge is the relationship of different terms and key phrases based on different operations and is represented in the form of nodes and edges. All these nodes and edges constitute a knowledge network called HESN. HESN is the combination of small networks, consisting of relationships among different terms, which are also tracked to know from which particular instruction and operation the small network has been formed. HESN is updated each time new information is learned. The methodology is suitable for extracting knowledge from iText. The current research was focused on iText found in industrial documents from the nuclear power plant domain. Limited test data were used to test the approach due to the confidentiality of information. Future work includes working with more data and more extensive domain knowledge, improved structure of HESN for better relations representation, and an information retrieval mechanism from HESN based on natural language query.

Author Contributions: H.A.G. and S.S.A.J.; methodology, H.A.G. and S.S.A.J.; software, S.S.A.J.; validation, H.A.G. and H.A.H.; formal analysis, H.A.G. and S.S.A.J.; investigation, H.A.G.; resources, H.A.H.; writing—original draft preparation, S.S.A.J.; writing—review and editing, H.A.G.; visualization, H.A.G. and S.S.A.J.; supervision, H.A.G. and J.R.; project administration, H.A.G.; funding acquisition, H.A.G. and J.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by NSERC and Ontario Power Generation (OPG).

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Not Applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhao, Y.; Smidts, C. A method for systematically developing the knowledge base of reactor operators in nuclear power plants to support cognitive modeling of operator performance. *Reliab. Eng. Syst. Saf.* **2019**, *186*, 64–77. [\[CrossRef\]](#)
2. Rodríguez-García, M.Á.; García-Sánchez, F.; Valencia-García, R. Knowledge-Based System for Crop Pests and Diseases Recognition. *Electronics* **2021**, *10*, 905. [\[CrossRef\]](#)
3. Skobelev, P.; Simonova, E.; Smirnov, S.; Budaev, D.; Voshchuk, G.; Morokov, A. Development of a Knowledge Base in the “Smart Farming” System for Agricultural Enterprise Management. *Procedia Comput. Sci.* **2019**, *150*, 154–161. [\[CrossRef\]](#)
4. Ritou, M.; Belkadi, F.; Yahouni, Z.; Cunha, C.D.; Laroche, F.; Furet, B. Knowledge-based multi-level aggregation for decision aid in the machining industry. *CIRP Ann.* **2019**, *68*, 475–478. [\[CrossRef\]](#)
5. Zhong, W.; Li, C.; Peng, X.; Wan, F.; An, X.; Tian, Z. A Knowledge Base System for Operation Optimization: Design and Implementation Practice for the Polyethylene Process. *Engineering* **2019**, *5*, 1041–1048. [\[CrossRef\]](#)
6. Li, T.; Chen, Z. An ontology-based learning approach for automatically classifying security requirements. *J. Syst. Softw.* **2020**, *165*, 110566. [\[CrossRef\]](#)
7. Wu, C.; Wu, P.; Wang, J.; Jiang, R.; Chen, M.; Wang, X. Ontological knowledge base for concrete bridge rehabilitation project management. *Autom. Constr.* **2021**, *121*, 103428. [\[CrossRef\]](#)
8. Sanfilippo, E.M.; Belkadi, F.; Bernard, A. Ontology-based knowledge representation for additive manufacturing. *Comput. Ind.* **2019**, *109*, 182–194. [\[CrossRef\]](#)
9. Wątróbski, J. Ontology learning methods from text—An extensive knowledge-based approach. *Procedia Comput. Sci.* **2020**, *176*, 3356–3368. [\[CrossRef\]](#)
10. Nadeau, D.; Sekine, S. A survey of named entity recognition and classification. *Linguisticae Investig. Int. J. Linguist. Lang. Resour.* **2007**, *30*, 3–26. [\[CrossRef\]](#)

11. Bach, N.; Badaskar, S. A Review of Relation Extraction. *Lit. Rev. Lang. Stat. II* **2007**, *2*, 1–15.
12. Martinez-Rodriguez, J.L.; Lopez-Arevalo, I.; Rios-Alvarado, A.B. OpenIE-based approach for Knowledge Graph construction from text. *Expert Syst. Appl.* **2018**, *113*, 339–355. [[CrossRef](#)]
13. Kim, T.; Yun, Y.; Kim, N. Deep Learning-Based Knowledge Graph Generation for COVID-19. *Sustainability* **2021**, *13*, 2276. [[CrossRef](#)]
14. Bekoulis, G.; Deleu, J.; Demeester, T.; Develder, C. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Syst. Appl.* **2018**, *114*, 34–45. [[CrossRef](#)]
15. Wang, Z.; Xu, S.; Zhu, L. Semantic relation extraction aware of N-gram features from unstructured biomedical text. *J. Biomed. Inform.* **2018**, *86*, 59–70. [[CrossRef](#)] [[PubMed](#)]
16. Nie, B.; Sun, S. Knowledge graph embedding via reasoning over entities, relations, and text. *Future Gener. Comput. Syst.* **2019**, *91*, 426–433. [[CrossRef](#)]
17. Xu, B.; Zhuge, H. The influence of semantic link network on the ability of question-answering system. *Future Gener. Comput. Syst.* **2020**, *108*, 1–14. [[CrossRef](#)]
18. Guo, A.; Tan, Z.; Zhao, X. Measuring Triplet Trustworthiness in Knowledge Graphs via Expanded Relation Detection. In *Knowledge Science, Engineering and Management*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 65–76. [[CrossRef](#)]
19. Xiao, S.; Song, M. A Text-Generated Method to Joint Extraction of Entities and Relations. *Appl. Sci.* **2019**, *9*, 3795. [[CrossRef](#)]
20. Spacy.io. Industrial-Strength Natural Language Processing in Python. Available online: <https://spacy.io/> (accessed on 18 April 2021).