Ontology-Based Methodology for Knowledge Acquisition from Groupware

Chukwudi Festus Uwasomba *, Yunli Lee †, Zaharin Yusoff and Teck Min Chin

Abstract: Groupware exist, and they contain expertise knowledge (explicit and tacit) that is primarily for solving problems, and it is collected on-the-job through virtual teams; such knowledge should be harvested. A system to acquire on-the-job knowledge of experts from groupware in view of the enrichment of intelligent agents has become one of the important technologies that is very much in demand in the field of knowledge technology, especially in this era of textual data explosion including due to the ever-increasing remote work culture. Before acquiring new knowledge from sentences in groupware into an existing ontology, it is vital to process the groupware discussions to recognise concepts (especially new ones), as well as to find the appropriate mappings between the said concepts and the destination ontology. There are several mapping procedures in the literature, but these have been formulated on the basis of mapping two or more independent ontologies using concept-similarities and it requires a significant amount of computation. With the goal of lowering computational complexities, identification difficulties, and complications of insertion (hooking) of a concept into an existing ontology, this paper proposes: (1) an ontology-based framework with changeable modules to harvest knowledge from groupware discussions; and (2) a facts enrichment approach (FEA) for the identification of new concepts and the insertion/hooking of new concepts from sentences into an existing ontology. This unique approach can be implemented on any platform of choice using current or newly constructed modules that can be constantly revised with enhanced sophistication or extensions. In general, textual data is taken and analysed in view of the creation of an ontology that can be utilised to power intelligent agents. The complete architecture of the framework is provided and the evaluation of the results reveal that the proposed methodology performs significantly better compared to the universally recommended thresholds as well as the existing works. Our technique shows a notable high improvement on the F1 score that measures precision and recall. In terms of future work, the study recommends the development of algorithms to fully automate the framework as well as for harvesting tacit knowledge from groupware.

Keywords: knowledge technology; ontology; knowledge acquisition; knowledge representation; virtual software development

1. Introduction

Groupware exist, and they contain expertise knowledge (explicit and tacit) that is primarily for solving problems and it is collected on-the-job through virtual teams. Such knowledge should be harvested. In recent times, people and corporations are driven to the use of groupware, which the internet offers, to support the ever-increasing remote work culture [1]. A recent survey revealed that over 92% of employees use groupware for virtual collaboration [2]. This communication system in today’s internet era has become a knowledge store of raw data which results in an excellent source for knowledge harvesting to improve intelligent agents. Among the possible resources for developing intelligent...
agents is the know-how about the domain in which they are to be used [3]. Virtual teams’
groupware contains knowledge and practices regarding their specific fields, but in a format
(audio, video, and free text) that intelligent agents cannot understand directly. This on-the-
job expertise knowledge in groupware is readily available and should be harvested and
represented in some ontology to support the development of service robots and intelligent
agents, amongst others.

Groupware has been defined differently by various scholars [4], but in all of the defi-
nitions, being computer-based and used as a collaboration tool are the common traits. As
such, this study defines groupware as a computer-enabled distributed environment that
facilitates cooperation and coordination among a group of individuals that are working
toward a common goal. It is mostly developed to make knowledge creation and sharing
faster and easier in most organizations [5]. It also facilitates explicit and tacit knowledge
sharing in the organizations, according to Baronian [6]. Due to the benefits that they pro-
vide, groupware sites have exploded in popularity recently. Every day, a new groupware
application is created and launched, each with its own set of new features, flexibility, and
good usability. SourceForge, MicroExchange, Huddle, Fuze, and Drupal are just a few
examples of groupware systems that have been tested to handle high communication
volumes over time [7,8].

In AI, or simply intelligent agents, ontology plays a focal role to provide a communi-
cation framework that facilitates the definition of common vocabularies for applications
and other independent semantic functions [9]. As a conceptualization in an explicit specifi-
cation, intelligent agents need to have some knowledge about the domain in which they
operate, and the knowledge is to be represented accurately in them. This has been instanti-
ated in chatbots, question and answering systems, knowledge graphs, decision support,
and expert systems [10]. Owing to the benefits that remote work culture has offered [11],
on-the-job knowledge and practices of experts can be harvested and represented in the
form of an ontology to enhance the development of AI products.

Several empirical studies that have been conducted recently have revealed difficulties
in processing and acquiring new knowledge from unstructured text [12–15]. This is espe-
cially evident from groupware discussions, considering the nature of conversations which
are not written in formal sentence patterns and contain grammatical and typographical
errors. There are also problems in finding the mappings between the harvested knowledge
and the appropriate locations in the destination ontology or existing ontology since the
concept is coming from sentences and has less information [16]. A few studies have indeed
presented mapping procedures [17], matching [18–20], and alignment [21,22], but these
studies are formulated based on combining two or more independent ontologies, and
mostly using concept similarities that require a significant amount of computation and
calculations. These are ontology-to-ontology specific, hence the literature remains scant on
a technique for the identification and insertion (hooking) of new concepts from sentences
into an existing ontology.

Within the context of this study, concept identification represents the process of new
knowledge recognition while the insertion or hooking describes the procedure for adding
the newly recognised knowledge into an existing ontology. Considering a precondition for
hooking, the recognition of new knowledge is a core step needed with regard to adding
new concepts into an existing ontology. This is to ensure that all the distinct elements
in the ontology completely comprehend one another without concept duplications and
inconsistencies. Furthermore, the representation nature of concepts in ontologies has led
to having different concept hierarchies for equal, similar, or equivalent concepts in the
same ontology, more especially when a new concept has less information, no attribute and
values, and relations in/out such as being harvested from sentences.

The main contributions of this paper are to propose the following:

(1) An ontology-based framework with changeable modules to harvest knowledge from
groupware discussions. The uniqueness of this framework lies in its five processing
phases and components; earlier closet framework [23] focuses on event extraction with
four processing phases, which are covered by the first three phases in our proposed framework. This novelty of our framework is the inclusion of an acquisition hub and a knowledge chamber, which are not present in earlier frameworks.

2) A facts enrichment approach (FEA) for the identification and hooking of new concepts from sentences into an existing ontology, taking into consideration the notions of equality, similarity, and equivalence of concepts. The novelty of the FEA lies within its ability to identify and insert/hook a concept with less information such as those coming from sentences into an existing ontology.

The remaining part of this paper is organized as follows: Section 2 examines the current research trends in this domain; Section 3 introduces the design of the framework; Section 4 presents evaluation approaches, processes, and results; Section 5 provides a discussion of results; and finally, the conclusion and future work are provided in Section 6.

2. Literature Survey

Many researchers have proposed acquisition efforts to harvest knowledge from free text. This section presents some of the recent research endeavors that are highly related to this research. There are three subsections namely: Section 2.1 describes efforts on knowledge representation using an ontology, Section 2.2 provides available frameworks for knowledge extraction, and Section 2.3 presents the state-of-the-art for ontology equality of concepts (which is necessary for the recognition of new concepts). This includes a table to compare this study with the most recent related research in this area, thus highlighting on the contributions of this study.

2.1. Ontology in Knowledge Representation

Murtazina and Avdeenko [24] presented an ontology web language (OWL) ontology that stores knowledge about cognitive functions and techniques for assessing them. This included knowledge about the qualitative features of psychometric tests and the collection of data, including instances of a screening tests. The ontology can also be utilised to determine the link between cognitive functions and brain activity patterns. Building a consistent knowledge model in the field of cognitive functions assessment, classifying ontology instances, and detecting implicit relationships between class instances are all difficulties that OWL ontologies address. Qi et al. [25] suggested an ontology-based representation of urban heat island mitigation strategies (UHIMSs), focusing on the link between mitigation approaches, performance measures, and urban environments. The conceptualization of terminologies, the formation of linkages, and the integration are the three phases that make up our representation. Ebrahimipour and Yacout [26] proposed a method of representing knowledge using ontology concepts. It uses a bond-graph model to generate an equipment function structure that is related to fault propagation at the part component level. The combination of OWL and resource description framework (RDF [27]) are used to transform human words into a computer-readable representation. Parsing analysis, semantic interpretation, and knowledge representation are the phases in the methodology. Abburn [28] suggested the use of natural language processing (NLP) to extract relevant information from semi-structured and structured heterogeneous documents, then used RDF to represent the extracted information in a homogeneous and machine-understandable format, and then mapped the RDF triples to the appropriate concepts in disaster management domain ontologies.

Brono et al. [29] presented an ontology-based system for storing cultural information that may be used to manage and adapt a robot’s interaction to the user’s habits and preferences. This framework is based on three components namely: (i) relevant concepts, individual-specific, and preferences; (ii) program for individual-specific knowledge; and (iii) computational network for acquiring the individual-specific propagating knowledge. In Diab et al. [30], an ontology-based framework to share knowledge between humans and robots was proposed. This framework consisted of an environment for knowledge standardization, sensory module, and evaluation-based analysis for objects situation.
Larentis et al. [31] proposed an ontology to represent the knowledge of educational assistance in non-communicable chronic diseases (NCDs). Its goal is to assist educational formalities and systems that are designed for preventing and monitoring NCDs. The ontology is specified via competence questions, Semantic Web Rule Language (SWRL) rules, and Semantic Protocol and RDF Query Language (SPARQL) and is implemented in Protégé 5.5.0 using OWL. There are 138 classes, 31 relations, 6 semantic rules, and 575 axioms in the current version of the ontology. Although all these studies have pointed towards representing knowledge using ontologies, there has been no indication of representing on-the-job knowledge and practices of virtual software development teams from groupware.

For groupware systems, Vieira et al. [32] suggested an ontology to formally define context. Physical, organizational, and interaction contexts are the three basic categories of context information. They also showed how this ontology may be used for context inference, offering tools for user communication that are based on the current context of each user. The definition of classes, properties, and instances of these classes formalized a domain. They utilised Protégé 3.11 to change the ontology and axioms and the Java Embedded Object Production System (JEOPS) inference machine to construct the rules for context reasoning. Vieira et al. did not provide a clear method on how the raw data in groupware were processed or how the extracted concepts are inserted into an existing ontology, and there is also limited information on how the evaluation was done. In this study, we are focused on two major contributions: firstly, an ontology-based framework with changeable modules to acquire knowledge from groupware discussion; and secondly, designing a technique for the identification and hooking of new concepts from sentences into an existing ontology, taking into consideration the notions of equality, similarity, and equivalence of concepts.

2.2. Framework for Knowledge Extraction

Kertkeidkachorn [33] proposed T2KG, an automatic knowledge graph (KG) creation framework for natural language texts. The framework used similarity and rule-based approaches to align predicates. Entity mapping, coreference resolution, triple extraction, triple integration, and predicate mapping are the main components of this framework. An F1 score was used to evaluate this framework and it uses plain texts as the dataset. In Milosevic [34], a framework to extract numerical and textual information from tables in clinical literature was proposed. It consisted of six phases which included detection of tables, functional and structural processing, semantic tagging, pragmatic processing, cell selection, and syntactic extraction. Plain texts from clinical publications were used in this research and evaluation was by precision, recall, and the F1-score. Wang et al. [35] provided a unified methodology for extracting base facts and temporal facts from textual web sources in their framework. Candidate gathering, pattern analysis, graph creation, and label propagation were all factors that were examined in this approach. The framework’s input data source was Wikipedia, and it was evaluated by precision.

A unique framework was developed in Chuanyan et al. [36] for automatically extracting the temporal knowledge of entity relationships. Different parameters were explored in this proposed framework. These included heuristic data training, bootstrapping, Markov Logic Networks, pattern generation, and pattern selection. Precision, recall, and F-score were used to evaluate the framework and the data source was the internet. In Kuzey and Weikum [37], a framework for extracting temporal facts and events from Wikipedia articles’ free text and semi-structured data was proposed. The framework constructs a temporal ontology from data. The framework’s input data source was Wikipedia and it was evaluated by precision. Mahmood [38] proposed a knowledge extraction framework using finite-state transducers (FSTs) to extract named entities. This goes through five stages: content gathering, tokenisation, PoS tagging, multiword detection, and NER. F1 score, precision, and recall were adopted for the evaluation.

Abebe [23] proposed an event extraction approach for developing an event-based collaborative knowledge management architecture. Four parameters were taken into account
in this framework: dataset classifiers, uniform data model normalizer, event-based collective knowledge generator, and query formulatior. The framework’s input data source was social media and it was evaluated using the F1 score. A forensic framework for recognizing, gathering, investigating, detailing, and reporting content from the dark web was proposed in Popov et al. [39]. Identity, spidering, accessibility, structure parameters, analysis, and preservation were all taken into account in the suggested framework. The study used web data and ex-ante evaluation was done to assess the framework. Masum [40] proposed an automatic knowledge extraction framework to extract the most relevant sections of interest from a corpus of COVID-19-related research articles. The key components of this framework included query expansion, data pre-processing, transformation, similarity calculation, information extraction, and similarity network. The framework used data from scholarly articles. A conceptual framework for knowledge extraction and visualization was proposed by Becheru and Popescu [41]. A social media-based learning environment known as eMUSE was examined in the proposed framework. The study’s data came from social media and social network analysis was employed to assess the framework. Even though all these empirical studies focus on designing a framework for knowledge extraction, none of these frameworks is constructed with changeable modules and to acquire practices and procedurals from groupware. The closet framework is the one that is suggested in Abebe et al. [23] that was mentioned earlier, which focuses on event extraction with four processing phases that are covered by the first three phases in our proposed framework. This novelty of our framework is the inclusion of an acquisition hub and a knowledge chamber which are not present in the earlier frameworks.

2.3. Ontology Equality of Concepts

The notion of equality of concepts is necessary for the recognition of new concepts, when comparing with an existing ontology. This will be a major contribution in this research. Ngom et al. [42] introduced a method for validating the addition of a new concept to an ontology. The method functions in three stages, by firstly locating the neighborhood of the concept (C) within the basic ontology (Ob) and storing their semantic-similarity values in a stack. The neighborhood denotes concepts in Ob that are more similar to C. Secondly, the semantic similarity between C and its neighborhood that was discovered in the first stage was evaluated in the general ontology (Og). Finally, the correlation between the values that were noted in the previous stage was assessed. The authors used the whole WordNet as a general ontology and a WordNet branch as a basic ontology. They utilised the edge semantic similarity measure. To establish equivalence relations among concepts, Yin et al. [43] proposed a new approach that was based on Classification with Word and CONtext-Similarity (CWCONS). The core idea behind CWCONS is to categorize ontology tree nodes into two types: classification nodes and concept nodes, both of which rely on the ontology’s tree structure. To analyse similarities, they employed the longest common substring (LCS) and Tversky’s similarity model.

Xue et al. [44] proposed a similarity measure to calculate the similarity value of two ontology entities/concepts, after which an optimal model for the ontology matching problem is built; then, an evolutionary algorithm-based fully automated matcher is provided to solve the ontology matching issues; and finally, to balance the workload on user and the impact of his activity, concept hierarchy graph-based reasoning methodologies were proposed. Oliveira and Pesquita [45] suggested ontology matching algorithms that are capable of locating compound mappings across diverse biomedical ontologies. This is akin to ternary mappings, for example, asserting that “aortic valve stenosis” (HP:0001650) is comparable to the intersection of an “aortic valve” (FMA:7236) and is “constricted” (PATO:0001847). To cope with the higher computing demands, the algorithms used search space filtering that was based on partial mappings between ontology pairings. The evaluation of the algorithms was done with precision. Priya and Kumar [46] presented a granular computing approach for mapping numerous existing ontologies into a single representative domain ontology. It is made up of four granular computing processes: association,
isolation, purification, and reduction, which can be used to unify a set of related nodes in ontologies. The approach accomplishes ontology mappings by going through two phases: similarity calculation and granular computing. The evaluation was based on ontologies for transportation and vehicles.

To improve the generalization performance of a mapping between two ontologies, Liu et al. [47] proposed HISDOM, a novel ontology mapping system. HISDOM compares ontologies that are based on a variety of characteristics such as concept names, attributes, instances, and structural similarities. It calculates ontology mapping similarity using a convolutional neural network. The Ontology Alignment Evaluation Initiative (OAEI) dataset was used in HISDOM’s experiments. Ernadote [48] proposed a method for ensuring that the two ontologies remain aligned in the face of reconciliation restrictions that were defined between them. These restrictions deal with semantic linkages which help for better grasping of overlapping in different perspectives. The technique extension is provided from the user’s perspective before being theoretically formalized to achieve the goal of establishing a solid basis for comprehending limits and restrictions in concrete applications.

Maree and Belkhatir [49] proposed combining domain-specific ontologies that were based on multiple external semantic resources to address the semantic heterogeneity challenge. The proposed approach is based on making aggregated judgements on the semantic correspondences between the entities/concepts of different ontologies using knowledge that is represented by multiple external resources. Another two difficulties they addressed in their suggested approach were: (i) identifying and dealing with inconsistencies of semantic relations between concepts in an ontology; and (ii) using an integrated statistical and semantic technique to address the issue of missing background knowledge in the exploited knowledge bases. An ontology merger technique that is based on semantic similarity between concepts is proposed by Zhen-Xing and Xing-Yan [50]. It converts ontology into a formal context before calculating the semantic similarity of the concepts that are contained within. It obtains the ontology after reduction and concept lattice development. To integrate heterogeneous tourist information for online trip planning, Huang and Bian [51] introduced an ontology-based approach and a formal concept analysis (FCA) technique. In accordance with their respective perspectives, two ontologies were developed, one for travelers and one for tourism information suppliers. The ontology for travelers is based on research in the tourism field. Using the FCA approach, the ontology for tourist information providers is created by merging heterogeneous web tourism information.

A summary of the most recent research in this domain taken from the above review, including a comparison with this proposed research, is given in Table 1.

| Authors/Year | Source of New Knowledge | Research Focus | Sampled Ontology | Techniques Used | Evaluation |
|-------------|------------------------|----------------|-----------------|----------------|------------|
| Ngom et al. [42] | Existing ontology | Adding a concept from one ontology to another ontology | WordNet | Similarity measure | Correlation among ontologies |
| Yin et al. [43] | Existing ontology | Merging two or more existing ontologies | WordNet | Classification with Word and CONText Similarity | LCS |
| Xue et al. [44] | Existing ontology | Matching two or more existing ontologies | WordNet | Similarity measure | Recall, Precision, F-measure |
| Oliveira and Pesquita [45] | Existing ontology | Matching two or more existing ontologies | Biomedical ontologies | Similarity measure | Recall, Precision, F-measure |
| Priya and Kumar [46] | Existing ontology | Mapping two or more existing ontologies | Transportation and vehicles ontologies | Granular computing | Use case |
| Liu et al. [47] | Existing ontology | Mapping two or more existing ontologies | OAEI dataset | Mapping similarity | Accuracy |
| Ernadote [48] | Existing ontology | Aligning two or more existing ontologies | Metamodel-based ontologies | NA | NA |
Several salient points may be drawn from Table 1:

- Most of the previous research efforts in this domain are similar in that all are looking for new knowledge to add into a destination ontology from another existing ontology.
- Within such efforts, there is no clear consensus on the notion of equality, similarity, and equivalence of concepts, which is a necessity for the recognition of new concepts from any given source to be compared with an existing ontology.
- The literature is also scant on a technique for the insertion/hooking of a newly recognized concept into an existing ontology.

In sum, the literature is indeed scant on a formalized technique for the identification and insertion/hooking of a new concept into an existing ontology from other existing ontologies, let alone when the source is from sentences (free text), and especially from groupware. The novelty of our approach is thus the discovery of new concepts from sentences (in groupware) using a proposed FEA approach for the recognition of new concepts and the insertion/hooking into an existing ontology.

3. Design

This section presents the proposed framework (Figure 1), one of the major contributions of this paper. The design is novel in comparison to its forerunners and the closest comparable framework that is accessible, according to literature, is as suggested in Abebe et al. [23]. The said framework, event collective knowledge (eCK), is aimed at events extraction from social media and other multimedia digital ecosystem in general, whereas our proposed framework is focused on the acquisition of on-the-job knowledge, procedurals, and practices from virtual teams’ groupware. The datasets classifiers, uniform data model normalizer, collective knowledge generator, and query formulator are the four processing phases of eCK, which are related to the first three phases in our proposed framework. The five stages of our proposed framework are Groupware-Chamber (GC), Cleansing-Chamber (CC), Harvesting-Hub (HH), Acquisition-Hub (AH), and Knowledge-Hub (KH), which allow textual datasets to be processed from its free (unstructured) state to a refined (structured) knowledge in the form of ontologies for enabling intelligent agents.

The AH and KH are included in our proposed framework because they serve as a knowledge acquisition and a repository of ontologies modules, respectively, and these features are not present in the eCK. The framework that is proposed in this paper is semi-automated with reusable and incremental modules that have been formalized.

In terms of operability, all the components within the framework function collaboratively to ensure that textual data is taken and analyzed in view of the creation of an ontology that can be utilized to power intelligent agents. The input channel for entering text into the framework is the GC. The entered text is cleansed at the CC and dispensed in
a format that the HH can understand. At the HH, a variety of strategies are employed to extract knowledge from the cleansed text on the basis of their respective meanings. The AH then assists in validating whether the collected knowledge is an existing one (i.e., it is already in the target ontology) or a new one. This is done via the processes of identification and hooking whereby the recognised new knowledge is placed within the target ontology in the KH. Below are detailed descriptions of how each of these components operates as well as their methodologies and internal structures.

Figure 1. The ontology-based framework for harvesting knowledge from groupware.

3.1. Groupware Chamber

In the framework, the Groupware Chamber deals with input text issues. It is the only component through which any raw data can get into the framework. This chamber is the communication system that is used by virtual software development teams. The term “virtual software development team” refers to a notion in which the team members that are developing software collaborate from multiple locations through a computer-aided environment. This implies that the team members who perform tasks, team-leaders who supervise tasks, and the managers who regulate the project may not be situated at the same worksite. The textual conversations of the team in the groupware form the raw data or input that were used in this study. Although the conversations could also be in the form of diagrams, video, or audio, this ontology-based framework is designed only for text, be it structured, unstructured, or semi-structured.

3.2. Cleansing Chamber

The framework has a component that is responsible for ensuring that the input text is cleansed into the required format. This is the core function of the Cleansing Chamber. Considering that groupware raw data is, in essence, a conversation in which the language structure is not very formal, this chamber is considered an important function. All mentions, special characters, time, phone numbers, and other words that are deemed unusable are erased during the cleansing process. To ensure that input texts are adequately cleansed, the system employs NeatText [52], a python application. The “pip install nexttext” command can be used to install NextText. The docx.describe function can be used to describe and map irrelevant and unsuitable material, while the docx.remove function can be used to remove the mapped text. A spell-checker is also integrated to ensure that grammatical
and typographical errors are corrected. Sentences and/or paragraphs are the outputs of the Cleansing Chamber. In this study, the total sample is divided into four subsamples and then each subsample would undergo the cleansing process (Figure 2). At the end, 7313 words, forming 491 sentences according to the lexical word counter, would remain.

![Figure 2. The Cleansing workflow.](image)

### 3.3. Harvesting Hub

The key processing component of the framework is the Harvesting Hub. It is made up of tasks that are related to natural language processing (NLP) and a scenarios base, the latter being a technique to harvest tacit knowledge from experts [53]. As this study is mainly about groupware text, the emphasis is currently only on the NLP task. A morpho-syntactic parser and a logico-semantic parser are included in the NLP task (Figure 3). Stanford stanza, a python natural language analysis tool [54], is used to perform morpho-syntactic analysis to generate a morpho-syntactic structure from the sentence in the system. The input sentence is split down into its constituent elements, the POS (part of speech) is detected for each word occurrence, and the relationships between the words and phrases (syntagmatic groups) are computed at this level (syntactic functions). The logico-semantic parser is used to ascertain the appropriate meaning representation to the sentence components (semantic-features), and the sentence predicate-argument structures (PAS) are constructed to represent the meaning structure. All verbs, verbal nouns, and so on are mapped as predicates, and the arguments bear the WH (where? what? who? etc.) notions of the input (logico-semantic relations). This is how the framework deciphers the meaning of the words in the phrase or sentences.

Figure 4 presents an example of a logico-semantic structure that is generated from the Harvesting Hub. It shows how concepts and relations are harvested from the input sentence “team will perform an API configuration after the deployment”. Firstly, this sentence is broken down into its constituent elements to produce the syntactic functions as shown under the morpho-syntactic structure in Figure 4. Secondly, all the stopwords that do not affect the meaning of the words are dropped. As a result, the words “will” + “an” + “the” were removed, leaving “team”, “perform”, “API”, “configuration”, “after”, “deployment”. Thirdly, semantic features are applied. During this process, “configuration” and “deployment” are converted into their root form, being “configure” and “deploy”. Thereupon, “team” is identified as agent, “perform” as relation, “API” as instrument, “configure”
as process/result, “after” as relation, and “deploy” as process/result. On the basis that this study focuses on process/result acquisition, the words ‘deploy’ and ‘configure’ are harvested as concepts because they are results, while the arguments and ‘after’ are the relations. There is no attribute found in the sentence. This output structure is referred to as the logico-semantic structure of the sentence and is the basis of the harvested knowledge. These harvested outputs will proceed to the AH for identification and hooking.

3.4. Acquisition Hub

The Acquisition Hub is the key and the most complex component within the framework as it is where knowledge is harvested. This stage is also referred to as the validation stage. Two fundamental activities take place at this hub:

(a) New knowledge is recognised.
(b) New knowledge is inserted/hooked into the existing knowledge base.

To recognise a new knowledge is to identify one of the following:
- an entirely new concept,
- an existing concept with a new relation,
- an existing concept with a new attribute.

This step may seem quite trivial, especially if done manually, but automating it would require a major formalisation of EQUALITY of two concepts to be able to say that a given concept already exists in the knowledge base. In general, the notion of ontology equality of concepts is still not well-defined and there is no clear consensus on this [42–51], especially when the new knowledge is from a sentence, which would obviously contain much less content compared to a concept already defined in an existing knowledge base. Therefore, this paper, in its specificity, proposes a technique known as Facts Enrichment Approach (FEA) for the identification and insertion/hooking of a new concept (C) from a sentence into an existing ontology (BO) by considering the notions of equality, similarity, and equivalence of concepts to develop a Target Ontology (TO). TO includes the structural representations in BO, and all its concepts and the newly added C. It is pertinent to note that new knowledge is to be recognised from a sentence and, as such, it is quite rare to be able to recognise new attributes, but perhaps new relations. For the moment, the focus will be mainly on recognising new concepts only.
3.4. Acquisition Hub
The Acquisition Hub is the key and the most complex component within the framework as it is where knowledge is harvested. This stage is also referred to as the validation stage. Two fundamental activities take place at this hub:

(a) New knowledge is recognised.

(b) New knowledge is inserted/hooked into the existing knowledge base.

To recognise a new knowledge is to identify one of the following:

- an entirely new concept, or
- an existing concept with a new relation, or
- an existing concept with a new attribute.

This step may seem quite trivial, especially if done manually, but automating it would require a major formalisation of EQUALITY of two concepts to be able to say that a given concept already exists in the knowledge base. In general, the notion of ontology equality of concepts is still not well-defined and there is no clear consensus on this [42–51], especially when the new knowledge is from a sentence, which would obviously contain much less content compared to a concept already defined in an existing knowledge base. Therefore, this paper, in its specificity, proposes a technique known as Facts Enrichment Approach (FEA) for the identification and insertion/hooking of a new concept (C) from a sentence into an existing ontology (BO) by considering the notions of equality, similarity, and equivalence of concepts to develop a Target Ontology (TO). TO includes the structural representations in BO, and all its concepts and the newly added C. It is pertinent to note that new knowledge is to be recognised from a sentence and, as such, it is quite rare to be able to recognise new attributes, but perhaps new relations. For the moment, the focus will be mainly on recognising new concepts only.

To better understand the technicalities underlying our proposition, the general structure of a concept is presented in Figure 5:

Figure 5. Structural view of a concept.

Where: C is the concept in question, $S_j$ is a super-concept of C, IS_A, $R_I$, and $R_m$ are relations; $a_{tr_i}$, $a_{tr_k}$ are attributes with values, $v_i$, $v_k$, respectively; and $A_l$ $B_m$ are concepts related, respectively, in and out of C.

Figure 5 says that in general, for a concept in question C, we have:

Label = C
If C has a set of attributes and values \( \text{Catr}_i = v_i \) for certain indices \( i \), we write:

\[
\left[ \text{C} \ \{ \ \text{Catr}_i, v_i \ \} \right]
\] (1)

Should C have one or more super-concepts \( S_j \), i.e., there are relations \( \text{IS}_A \) (C, \( S_j \)) for certain indices \( j \), and with their respective attributes and values \( S_{j\text{atr}}_k = v_k \) for certain indices \( k \), we thus have:

\[
\left[ S_j \ \{ \ S_{j\text{atr}}_k, v_k \ \} \right]
\] (2)

As such, given inheritance, the set of attributes for the concept C is actually:

\[
\left[ \text{C} \ \{ \ \text{Catr}_i, v_i \ \} \cup \sum_j \left\{ S_{j\text{atr}}_k, v_k \right\} \right]
\] (3)

There may be one or more relations \( R_l \) into C from concepts \( A_l \), for certain indices \( l \), say:

\[
R_l \ (A_l, C)
\] (4)

There may be one or more relations \( R_m \) out of C into concepts \( B_m \), for certain indices \( m \), say:

\[
R_m \ (C, B_m)
\] (5)

Our proposed FEA uses relative meanings, a five-dimensional (5D) approach and certain fundamental principles (FP) to determine the newness of a concept, hence this is quite different from the approaches that are in \([41,43,44,49]\), which rely on similarity measures that may vary sometimes. Our approach checks all the labels, attributes and values, relations (in/out), and associated super-concepts and sub-concepts. The FP that guides the identification of a new concept using FEA are:

- Labels need not be the same at the outset but should be made the same once recognised as equal, similar and/or of equivalence.
- The set of attributes and values cannot be expected to be the same, but they must not contradict. Once they are deemed to be the same, then the attributes must be unioned (take the union of both sets).
- The relations in and out of the concept cannot be expected to be the same, but they must not contradict. Once they are deemed to be the same, then the relations also must be unioned.

Examples of cases where there is No Contradiction within attributes and relations in/out include the following:

- one attribute in a concept labelled B but not in a concept labelled C.
- the same for relations in and out of B and C.
- but if the same attribute is in both, the values cannot be different,
- if the same relations exist, they must go to or come from the same concepts.

In more formal terms for the last two cases, we have (for certain indices \( I, m, n, x, y \)):

- if the same attribute is in both, the values cannot be different,
  if \( (\text{Batr}_i, V_m) \) in B and \( (\text{Catr}_i, V_n) \) in C

\[
\text{Then, } V_m = V_n
\] (6)

- if the same relations exist, they must go to or come from the same concepts
  if both \( R_x \) (\( A_m, B \)) and \( R_x \) (\( A_n, C \)) exist

\[
\text{Then, } A_m = A_n
\] (7)

if both \( R_y \) (\( B, A_m \)) and \( R_y \) (\( C, A_n \)) exist

\[
\text{Then, } A_m = A_n
\] (8)
The techniques that are proposed in this paper are formed on the basis of ontology equality of concepts to determine if a given concept is equal to an existing concept in the knowledge base, hence not a new concept. The notion of equality is very much related to the notions of equivalence and similarity, which we outline below. First, recall that a concept is basically made up of a label, attributes with values (including inherited ones), and relations in and out with other concepts.

Equality

- Exact equality can be when all are the same (label, attributes and values, and relations in/out). This is, however, not very likely to be obtained since the inputs are sentences, hence with lesser information. Other forms of equality may be defined in terms of some level of equivalence and/or similarity.

Equivalence

- Equivalence characterises a condition of being equal or equivalent in value, worth, function, etc. (e.g., equivalent equations are algebraic equations that have identical solutions or roots). This may translate to having different labels but with attributes and values and relations in/out that may be similar. Considering the aim of this study, this notion is not very useful at the moment.

Similarity

- Similarity describes having resemblance in appearance, character, or quantity without being identical (e.g., similar triangles are the same shape <same angles>, but not necessarily the same size). It may be seen as being equal in certain parts but not in all, with these being defined at the level of label, attributes, and relations in/out. This is the most useful but there must be no contradictions in the parts with some commonality.

In general, given two concepts, each with their respective three aspects (label, attributes and values, relations in and out), they are deemed to be:

- Exactly equal if all the three aspects are exactly the same. If this is obtained, then it is not a new concept.
- May or may not be equal if some aspects belong to one but not the other (and vice versa).
- Not equal if there are any contradictions within the attributes and values and/or relations in and out (excluding labels). If this is obtained, then it may be equal to another concept.

Some example situations include the following:

- May still be equal, as the labels may be synonyms, or in different languages, and so on.
- May still be equal, as they differ in labels but no contradictions in the attributes.
- Not equal, as there is a contradiction in the attribute.
- May still be equal, as they differ in labels but no contradictions in the relations in and out.
- Not equal, as there is a contradiction in the relations in and out.

The following notion of equality is heuristically driven and is currently adopted, but is best checked manually (all three need to be satisfied):

- The labels are the same or synonymous.
- No contradictions.
- Have at least one or more exactly same attributes or relations in/out.

Not equal to any existing concept means it is a new concept, in other words a newly harvested concept.

We now look at the insertion/hooking of a newly harvested concept into the knowledge base. This is still a preliminary study as hooking will be worked on for a better formalization and will be reported in a later publication.

A newly harvested concept (C) comes from a sentence that is then transformed into a logico-semantic structure, which is essentially in the form of a basic ontology, thus having a
label and potentially attributes with values and relations in and out. However, as it comes from a sentence, attributes and values are rarely found, but perhaps relations in and/or out. With relations in and/or out, say $R_l(A_l, C)$ and/or $R_m(C, B_m)$ (see Equations (4) and (5)), then C can be readily hooked to the concepts $A_l$ and/or $B_m$ (provided, of course, that these latter concepts are existing in the knowledge base).

It is when there are no such relations that difficulties would arise. An illustration is given in Figure 6 where the logico-semantic structure is the shaded part in the bottom left-hand corner and the newly harvested concept is ‘api’. The righthand side of the figure is part of the existing ontology and the challenge is to discover the following relation, which is the hook:

$$\text{Is-A}(\text{api}, \text{ArtifactDev})$$

Once found, then the picture would be complete. If the source had been from another knowledge base, together with attributes and values, then the Is-A relations can be discovered by comparing the sets of attributes where the set of attributes and values of the super-concept would be a subset of the set of attributes and values of the sub-concept. With the source being sentences, this part remains a challenge. For the moment, this is done manually.

The sequence of actions that happens during the identification and hooking process is presented in Figure 7. It involves checking decision nodes in the ontology hierarchical structure and can only take place once there are no contradictions, as explained above. As stated earlier, the workflow focuses on new concepts only and the textual data for this study is software-development-based. Firstly, it starts with having a harvested concept (C), from a sentence, then it passes through the FEA mechanism, where all the components of C (label, attributes and values, relations in/out) are checked to ascertain if there are similar or synonymous concepts in the ontology. The study adopts the software development process handbook of IEEE (SWEBOK), PMBOK, and RUP as the reference base (RB) for
synonyms for a new label. In so doing, the neighborhood of C is determined in the existing ontology. For the label, if it is equal to one already existing, it is made similar and thereafter, unioned, but, if it is not equal, it is recognised as a new concept. For attributes and values and relations in/out, if there is similarity, it is unioned, but if it is not similar, a common relation is determined. Once the common relation is found whether in/out, C is recognised as new knowledge. Secondly, the recognised new knowledge is hooked. The hooking process is looped to ensure that the new knowledge is hooked or inserted at an appropriate position depending on the available relation in/out. It is either hooked as a sibling of a super-concept, sub-concept, child, or offspring using Is-A or as an argument (Figure 6). Emphases are placed mostly on Is-A and argument relations in this study.

**Figure 7.** The hooking workflow.
3.5. Knowledge Hub

The framework that is described in this paper is ontology-based as all the output is stored in the form of an ontology in the Knowledge Hub. This makes the Knowledge Hub a key component of the framework. Given that the demonstration for this study is centered on groupware raw data from a virtual software development team, an ontology was created to accommodate the concepts and sub-concepts in this domain as there is little literature on the subject. This hub was built using Protégé, a renowned ontological engineering tool [55], with the five top-level ontologies being: SoftwareType, ProjectTools, CommonProjectIssues, ProjectIssueSolution, and WagileElement serving as the foundation. Top-level ontologies such as CommonProjectIssues, ProjectIssueSolution, and ProjectTools were chosen because they represent the core process areas in the software engineering process [56]. WagileElement was built on the footing of the rational unified process framework [57], waterfall, and agile process models [58]. The SoftwareType was added as a top-level ontology by the experts during the confirmation and acquisition of more knowledge. The Knowledge Hub’s internal structure is depicted in Figure 8.

![Figure 8. Internal structure of the knowledge hub.](image)

From Figure 8, SoftwareType, as the names suggest, is created to validate and store (if new) all the software types harvested from the groupware. ProjectTools, as a top-level ontology, is designed to verify and place all the tools and platforms that are used by the virtual software development team that are acquired from the groupware discussion. CommonProjectIssues is adopted as a top-level ontology to identify and store all the common project issues that are harvested from the groupware. ProjectIssueSolutions, as a top-level ontology, is designed to validate and place all the solutions to the identified issues in the virtual software development team groupware discussions, while WagileElement is designed to validate and store the disciplines, artifacts, tasks, and roles that are identified in the discussions of the virtual software development team. Within the context of this paper, WagileElement is regarded as the element of the hybrid software development process denoting the combination of both waterfall and agile methodologies. These elements
were used in the development of the Knowledge Hub to effectively define the relationships among the discipline, artifacts, tasks, and roles within the team. The discipline concept is linked to the fundamental key knowledge areas in the software development environment (Table 2).

Table 2. The relationship between discipline concept and wagile roles.

| Discipline Concept             | Waterfall Role | Agile Role | WAGILE Role |
|--------------------------------|----------------|------------|-------------|
| Project-management            | PM             | PO         | PM          |
| Team-focus                    | PM             | SM         | SM          |
| Product-ownership             | PM             | PO         | PO          |
| Requirement-analysis          | BSA            | DT         | BSA         |
| Design                        | PM & UX        | DT         | UX          |
| Implementation                | Dev            | DT         | Dev         |
| Test/QA                       | QA             | DT         | QA          |
| Deployment                    | RM             | DT         | RM          |
| Maintenance                   | ASA            | DT         | ASA         |

PM = Project Manager, BSA = Business System Analyst, UX = UX Designer, Dev = Developers, QA = Quality Assurance/Testers, RM = Release Manager, ASA = Application support Analyst, PO = product owner, SM = Scrum Master, DT = Development Team.

The WagileElement enables us to aggregate all areas and roles that are needed for this study. It underscores discipline with respect to its main components, which place emphasis on roles (the performer of the actions that produce input/output artifacts), tasks (how the actions are performed), and artifacts (what the actions produce). The role characterizes the behaviour, attitude, and duties of an individual or group of people working as a team. It stipulates the overall description of actions and artifacts that the role is responsible for. A task is the step-by-step activities for a piece of work that are performed as a result of the role. It describes the role that is accountable for the action and the artifacts that are required as input as well as the corresponding output. An artifact is generally regarded as a document, element, or model that is produced or utilized by the process, and it also records the role that is responsible for the artifact, as well as some other elements such as guidelines, templates, whitepapers, reports, and legacy tools that supplement the discipline components that were mentioned earlier.

In terms of operability within the framework, each of these top-level ontologies already consists of concepts with a label, attributes and values, and relations in/out. Therefore, a new concept from the AH is checked at the KH to determine if it already exists or not. If it does not, it is then identified as new knowledge. Only new knowledge will be hooked into the ontology. In so doing, the destination ontology is updated with current trends and innovations in the domain which, in turn, will provide a significant background when integrated into AI.

Table 2 presents the relationship between discipline concept and wagile roles used in this study. The table shows how software development roles in waterfall and agile which belong to a knowledge area in the discipline concept were aggregated to form wagile roles. The waterfall approach has seven roles (PM, BSA, UX, Dev, QA, RM, and ASA) while the agile approach has three roles (PO, SM, and DT). Wagile roles is the representation of all the core roles found in both waterfall and agile approaches. These are PM, BSA, UX, Dev, QA, RM, ASA, PO, SM, and DT.

A total of five different ontologies were developed in this study namely, Basic Ontology, Base Ontology, Target Ontology, Grand Ontology 1, and Grand Ontology 2. The entire methodology and strategy for creating these ontologies are presented in Figure 9. It starts with the creation of an initial ontology, which is then confirmed through expert consultation and the acquisition of further knowledge. Externalization, a concept that defines making tacit knowledge within experts explicit, is permitted at this level. Following that, concepts are formalized, and a Base Ontology (BO) is created at the end of the process. Since the framework is meant to harvest groupware raw data, experts only monitored the manual
augmentation of all inputs, concepts, attributes, and relations in the 491 sentences in the cleansing chamber into the BO to build the Target Ontology (TO), which is then taken as the ground truth in this study. Externalization is not permitted throughout this phase. This is to ensure that only the inputs from the Cleansing Chamber are used. To develop Grand Ontology 1, 20 sentences from the cleansing chamber are used. Facts were harvested at the Harvesting Hub, followed by validation (identification and hooking) in the Acquisition Hub, and then the label is added in the BO. This process is repeated in the development of Grand Ontology 2 but with 10 sentences from the 491 sentences at the cleansing chamber, but outside of the 20 that are already used.

Figure 9. The methodology and strategy for creating the ontologies.

4. Evaluation

This section presents the evaluation approaches and processes that are used in evaluating the framework and the results. At the moment, the evaluation focuses on the outputs of the framework, namely the ontologies. Given that this is a new area of study under knowledge acquisition, the evaluation was carried out between internal ontologies and existing works. Firstly, the ontology that is manually augmented with all concepts in the cleansed text from groupware which is referred to as target ontology (TO) was compared against two other ontologies that were developed using AH to harvest their concepts. These are Grand Ontology 1 (GO1), which is an ontology developed with randomly selected 20 sentences from the cleansed text, and Grand ontology 2 (GO2), another ontology that was developed with randomly selected 10 sentences from the cleansed text but outside of the 20 that were already used making a total of 30 sentences. Finally, a comparison was made with other existing works in this research direction.

4.1. Approaches

In this sub-section, the evaluation approaches that were used in evaluating the framework are discussed. There may have been some concept imbalance (erroneous hooking of concept or sub-concepts) during the validation phase, leading to limitations in accuracy. As such, a gold standard-based approach [59] was used. This approach is mostly used to compare between learned ontology and a predefined ontology, referred to as “a gold standard” [60]. The approach becomes very useful as this study compares two learned on-
ologies (GO1 and GO2) against a predefined ontology (TO). To achieve this, the confusion matrix \[61\] is utilized to evaluate error rate, accuracy, and F1-score.

The error rate was evaluated first with the assumption that errors occurred in hooking of new concepts from sentences into an existing ontology using FEA. Secondly, the accuracy of the identification and hooking were measured to evaluate the accuracy performance. Finally, the F1-score was evaluated to measure the precision and recall of the hooking concept. The classification metrics were defined according to the threshold of Ruuska et al. \[62\] as follows:

- **Error rate (ERR)** (1) is calculated as the number of all incorrect predictions divided by the total number in the dataset. The acceptance rate should be less than 0.10 indicating a minimal error.

\[
ERR = \frac{FP + FN}{TP + TN + FP + FN} \tag{9}
\]

- **Accuracy (ACY)** (2) is calculated as the number of all of the correct predictions divided by the total number of the dataset. The acceptance rate should be greater than 0.90 showing excellent classification.

\[
ACY = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}
\]

- **F1-Score (3)** is a harmonic mean of precision and recall. The acceptance rate should be greater than 0.90 suggesting excellent precision and recall.

\[
F1 - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}
\]

For comparison with existing works, we compared our results with Xue et al. \[44\], Oliveira and Pesquita \[45\], and Maree and Belkhatir \[49\]. This is on the basis that these recent studies are within the domain of this research and computed the same evaluation measures.

### 4.2. Processes

This sub-section goes into the evaluation processes that were used in this research. The harvested concepts from GO1 and GO2 were compared against the TO. With this, all 30 sentences were used for the evaluation. It was carried out manually by 40 volunteers that were enlisted to test the framework, and their responses were recorded using a confusion matrix. Before the evaluation began, the evaluators received a basic explanation of the functionalities of the components of the framework, were made to understand the purpose of the evaluation, and were knowledgeable of the expectations. Since the identification and hooking that were performed in this study focuses on concepts, the evaluators were instructed to rate if:

1. Concepts were properly named, and
2. Their hooking locations in the ontology were accurate.

The list of harvested concepts that were used in this evaluation is presented in Table A1, Appendix A. It consists of 30 concepts which were used for rating the concept naming, and then their hooking locations by the 40 evaluators. This resulted in 60 classification ratings for each evaluator which amounted to 2400 expected classifications. However, some ratings were incomplete, in all only 2126 classifications were received from the 40 evaluators (Table 3), and thus the analysis was based on those.

### 4.3. Results

This sub-section provides the results of the evaluation. Table 3 presents the summary of the classification by the evaluators. From the table, a total of 2126 responses were received from the 40 evaluators. GO1 accounted for 1449 responses, which translates to 724 and 725 responses for concept naming and hooking locations, respectively, while GO2
accounted for 667 that translates to 339 and 338 responses for concept naming and hooking locations, respectively.

Table 3. Summary of classification by the evaluators.

| Matrix Variables | Grand Ontology 1 | Grand Ontology 2 | Total |
|------------------|-----------------|-----------------|-------|
|                  | Concept Naming  | Hooking Location| Concept Naming | Hooking Location |       |
| True-positive    | 457             | 456             | 213 | 214 | 1340 |
| False-positive   | 32              | 34              | 15   | 14  | 95 |
| True-negative    | 232             | 231             | 109  | 106 | 678 |
| False-negative   | 3               | 4               | 2     | 4   | 13 |
| Total            | 724             | 725             | 339  | 338 | 2126 |

The computed results of the confusion matrix are presented in Table 4. The table shows an error rate of 0.05, an accuracy rate of 0.95, and an F1 score of 0.96 for concept naming in GO1; and an error rate of 0.05, an accuracy rate of 0.95, and an F1 score of 0.98 for concept naming in GO2. For hooking locations, an error rate of 0.05 is recorded for GO1 and GO2, an accuracy rate of 0.95 is also recorded for both ontologies, as well as an F1 score of 0.96.

Table 4. The computed of the confusion matrix.

| Parameters     | Grand Ontology 1 | Grand Ontology 2 |       |
|----------------|-----------------|-----------------|-------|
|                | ER   | ACY   | F1    | ER   | ACY   | F1    |
| Concept Naming | 0.048 | 0.952 | 0.963 | 0.050 | 0.949 | 0.984 |
| Hooking locations | 0.052 | 0.948 | 0.960 | 0.053 | 0.947 | 0.957 |
| Mean           | 0.05 | 0.95 | 0.96 | 0.05 | 0.95 | 0.97 |

Table 5 presents the comparative results between this research, the threshold of Ruuska et al., 2018, and the selected existing works as stated earlier. This comparison analysis is based on the F1 score which measures the precision and recall because that is the only matrix variable that was measured in this research and in existing works. Comparing our research to the thresholds, the table indicated an F1 score of 0.97 for this research where the recommended threshold is greater than 0.90 for high precision and recall. The threshold further recommends greater than 0.90 for accuracy, which is for correct identification and hooking, and this research recorded 0.95. The acceptable error rate as placed by the threshold is less than 0.10, while our classification logged a 0.05 error rate. In comparison with existing works, the table also revealed 0.81, 0.96, and 1.0 for Xue et al. [44], Oliveira and Pesquita [45], and Maree and Belkhatir [49], respectively, thus showing more favorable results for our approach.

Table 5. Comparative results between this research and existing studies.

| Matrix Variable | Thresholds | This Research | Xue et al. [43] | Oliveira and Pesquita [44] | Maree and Belkhatir [48] |
|-----------------|------------|--------------|-----------------|--------------------------|------------------------|
| F1 score        | greater than 0.90 | 0.97 | 0.81 | 0.96 | 1.0 |
| Accuracy        | greater than 0.90 | 0.95 | -   | -   | -   |
| Error rate      | less than 0.10 | 0.05 | -   | -   | -   |

5. Discussion

This study presented a technique for the identification and hooking of concepts from sentences into an existing ontology. It provides a novel approach for on-the-job expertise knowledge in groupware to be harvested and represented in ontology to support the development of service robots and intelligent agents. In terms of evaluation, the F1 score of 0.97 means that the technique has higher precision and recall and is very much in line with
the universally recommended thresholds. This suggests that our technique can be used to update an existing ontology with current trends and innovations coming from sentences. This, in turn, will provide significant background for AI development and deployment. The results further indicate that the error rate (0.05) and accuracy rate (0.95) are well aligned to the suggestions of the universal thresholds (less than 0.10 and 0.90, respectively). This is attributed to the accurate naming and hooking of the harvested new concepts vis-a-vis the destination ontology. Comparison with other existing works also holds that the proposed technique can be deemed to be a success. This is based on the comparative analysis of the F1 scores, which shows that the proposed technique is significantly better than Xue et al. [44], similarly for Oliveira and Pesquita [45], but performs slightly better than Maree and Belkhatir [49].

Getting recent literature with the same evaluation scope to perform a comparative analysis is rather challenging, and, as such, this was based on the F1-scores only. Given the situation, the error rate and accuracy rate were compared only with the threshold, and not yet with other existing works. As literature is scant on mapping concepts from sentences into an existing ontology, comparisons were made with results of automated mapping and merging of concepts between ontologies even though ours were carried out manually in this study. In the future, we plan to fully automate the processes in this technique and then re-evaluate with more comparable balanced matrix variables.

6. Conclusions & Future Work

In this paper, we presented a novel ontology-based framework for knowledge acquisition from sentences (text) in groupware, as well as a technique for the identification and hooking/insertion of new concepts into an existing ontology. The framework is currently semi-automated and has five main components, to take in textual data, analyze it, and to update an existing ontology that can be utilized to power intelligent agents. Ontology plays a focal role to provide a communication framework that facilitates the definition of common vocabularies in AI applications. However, most of the previous research efforts in this domain are similar in that all are looking for new knowledge to add into a destination ontology from another existing ontology. Within such efforts, there is no clear consensus on the notion of equality, similarity, and equivalence of concepts, which is a necessity for the recognition of new concepts from any given source to be compared with an existing ontology. In addition, the literature is indeed scant on a formalized technique for the identification and insertion/hooking of a new concept into an existing ontology from other existing ontologies, let alone when the source is from sentences (free text), and especially from groupware. The novelty of our approach is thus the discovery of new concepts from sentences (in groupware) using a proposed FEA approach for the recognition of new concepts and the insertion/hooking of the new concepts into an existing ontology.

In terms of evaluation, the F1 score of 0.97 means that the technique has higher precision and recall which is very much in line with the universal recommendations for the threshold. The results further indicate that the error rate (0.05) and accuracy rate (0.95) are also aligned with the suggestions for the universal thresholds (less than 0.10 and 0.90, respectively). Comparison with other existing works also holds that the proposed technique can be deemed to be a success. This is based on the comparative analysis of the F1 scores that show that the proposed technique is slightly better than those that are available in the literature. As literature is scant on mapping concepts from sentences into an existing ontology, comparisons were made with results of automated mapping and merging of concepts between ontologies even though it was done manually in this study. In the future, we plan to fully automate the framework and all its processes and then re-evaluate with more comparable balanced matrix variables, most especially using datasets from electronic clinical records and agricultural practices domains. We also plan to incorporate the acquisition of tacit knowledge from groupware.
Author Contributions: Conceptualization, C.F.U.; Methodology, C.F.U.; Supervision, Y.L., Z.Y. and T.M.C.; Writing—original draft, C.F.U.; Writing—review and editing, Y.L. and Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: Sunway University Postgraduate Research Support. This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of Sunway University, Malaysia.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

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

Abbreviations

The abbreviations used in this paper are as follows:

List of Abbreviations

| Abbreviation | Description |
|--------------|-------------|
| AH           | Acquisition Hub |
| AI           | Artificial Intelligence |
| CC           | Cleansing Chamber |
| FEA          | Facts Enrichment Approach |
| FCA          | Formal Concept Analysis |
| FN           | False negative |
| FP           | False Positive |
| GC           | Groupware Chamber |
| GO1          | Grand Ontology 1 |
| GO2          | Grand Ontology 2 |
| HH           | Harvesting Hub |
| II           | Identification Instrument |
| JEOPS        | Java Embedded Object Production System |
| KG           | Knowledge Graph |
| KH           | Knowledge Hub |
| LCS          | Longest Common Substring |
| OAEI         | Ontology Alignment Evaluation Initiative |
| OWL          | Web Ontology Language |
| RDF          | Resource Descriptive Framework |
| SWRL         | Semantic Web Rule Language |
| TN           | True Negative |
| TO           | Target Ontology |
| TP           | True positive |

List of Mathematical Symbols

| Symbol | Description |
|--------|-------------|
| C      | the concept in question. |
| S_j    | a super-concept of C. |
| IS_A, R_l, R_m | relations. |
| atr_k, atr_k | attributes with values. |
| v_i, v_k | attributes values. |
| A_l, B_m | related concepts in and out of C. |
Appendix A

Table A1. Testing datasets.

| Grand Ontology 1 | Grand Ontology 2 |
|------------------|------------------|
| Sentences Concepts | Sentences Concepts |
| 1 api            | 1 gitflow        |
| 2 code(new)      | 2 conversation   |
| 3 vertical scale | 3 milestone      |
| 4 cookie         | 4 -              |
| 5 feature(new)   | 5 bug            |
| 6 testlog        | 6 impact         |
| 7 signoff        | 7 journey        |
| 8 rollback       | 8 misuse         |
| 9 script         | 9 toggle         |
| 10 production    | 10 debt          |
| 11 go-no-go      | -                |
| 12 attend        | -                |
| 13 codebase      | -                |
| 14 build         | -                |
| 15 codefreeze    | -                |
| 16 timelines     | -                |
| 17 sprint        | -                |
| 18 backlog       | -                |
| 19 tag           | -                |
| 20 log           | -                |

References

1. Wang, B.; Liu, Y.; Qian, J.; Parker, S.K. Achieving Effective Remote Working during the COVID-19 Pandemic: A Work Design Perspective. *Appl. Psychol.* 2021, 70, 16–59. [CrossRef] [PubMed]
2. Sako, M. From remote work to working from anywhere. *Commun. ACM* 2021, 64, 20–22. [CrossRef]
3. Saba, D.; Sahli, Y.; Maouedj, R.; Hadidi, A.; Medjahed, M.B. Towards Artificial Intelligence: Concepts, Applications, and Innovations. In *Enabling AI Applications in Data Science*; Springer: Cham, Switzerland, 2021; pp. 103–146. [CrossRef]
4. Gacitua, R.; Astudillo, H.; Hitpass, B.; Osorio-Sanabria, M.; Taramasco, C. Recent Models for Collaborative E-Government Processes: A Survey. *IEEE Access* 2021, 9, 19602–19618. [CrossRef]
5. Xanthopoulou, S.; Kessopoulou, E.; Tsiotras, G. KM tools alignment with KM processes: The case study of the Greek public sector. *Knowl. Manag. Res. Pract.* 2021, 19, 1–11. [CrossRef]
6. Kodama, M. *Managing IT for Innovation: Dynamic Capabilities and Competitive Advantage*; Routledge: London, UK, 2021; ISBN 9780367462987.
7. Uwasomba, C.F.; Seeam, P.; Bellekens, X.; Seeam, A. Managing knowledge flows in Mauritian multinational corporations: Empirical analysis using the SECI model. In Proceedings of the 2016 IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), Balclava, Mauritius, 3–6 August 2016; pp. 341–344. [CrossRef]
8. Confalonieri, R.; Weyde, T.; Besold, T.R.; Martin, F.M.D.P. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artif. Intell.* 2021, 296, 103471. [CrossRef]
9. Rodrigo, A.; Peñas, A. A study about the future evaluation of Question-Answering systems. *Knowl.-Based Syst.* 2017, 137, 83–93. [CrossRef]
10. Charalampous, M.; Grant, C.A.; Tramontano, C.; Michailidis, E. Systematically reviewing remote e-workers’ well-being at work: A multidimensional approach. *Eur. J. Work Organ. Psychol.* 2019, 28, 51–73. [CrossRef]
11. Juric, D.; Stoilos, G.; Melo, A.; Moore, J.; Khodadadi, M. A System for Medical Information Extraction and Verification from Unstructured Text. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13314–13319. [CrossRef]
12. Ye, X.; Lu, Y. Automatic Extraction of Engineering Rules from Unstructured Text: A Natural Language Processing Approach. *J. Comput. Inf. Sci. Eng.* 2020, 20, 034501. [CrossRef]
13. Kolesnikov, A.; Kikin, P.; Niko, G.; Komissarova, E. Natural language processing systems for data extraction and mapping on the basis of unstructured text blocks. *Proc. Int. Conf. InterCarto/InterGIS* 2020, 26, 375–384. [CrossRef]
42. Ngom, A.N.; Diallo, P.F.; Kamara-Sangare, F.; Lo, M. A method to validate the insertion of a new concept in an ontology. In Proceedings of the 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Naples, Italy, 28 November–1 December 2016; pp. 275–281. [CrossRef]

43. Yin, C.; Gu, J.; Hou, Z. An ontology mapping approach based on classification with word and context similarity. In Proceedings of the 2016 12th International Conference on Semantics, Knowledge and Grids (SKG), Beijing, China, 15–17 August 2016; pp. 69–75. [CrossRef]

44. Xue, X.; Chen, J.; Ren, A. Interactive Ontology Matching Based on Evolutionary Algorithm. In Proceedings of the 2019 15th International Conference on Computational Intelligence and Security (CIS), Macao, China, 13–16 December 2019; pp. 1–5. [CrossRef]

45. Oliveira, D.; Pesquita, C. Improving the interoperability of biomedical ontologies with compound alignments. J. Biomed. Semant. 2018, 9, 13. [CrossRef]

46. Priya, M.; Kumar, C.A. An approach to merge domain ontologies using granular computing. Granul. Comput. 2019, 6, 69–94. [CrossRef]

47. Liu, J.; Tang, Y.; Xu, X. HISDOM: A Hybrid Ontology Mapping System based on Convolutional Neural Network and Dynamic Weight. In Proceedings of the 6th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, Auckland, New Zealand, 2–5 December 2019; pp. 67–70. [CrossRef]

48. Ernadote, D. Ontology reconciliation for system engineering. In Proceedings of the 2016 IEEE International Symposium on Systems Engineering (ISSE), Edinburgh, Scotland, 4–5 October 2016; pp. 1–8. [CrossRef]

49. Maree, M.; Belkhadir, M. Addressing semantic heterogeneity through multiple knowledge base assisted merging of domain-specific ontologies. Knowl.-Based Syst. 2015, 73, 199–211. [CrossRef]

50. Zhen-Xing, W.; Xing-Yan, T. Research of Ontology Merging Based on Concept Similarity. In Proceedings of the 2015 Seventh International Conference on Measuring Technology and Mechatronics Automation, Nanchang, China, 13–14 June 2015; pp. 831–834. [CrossRef]

51. Huang, Y.; Bian, L. Using ontologies and formal concept analysis to integrate heterogeneous tourism information. IEEE Trans. Emerg. Top. Comput. 2015, 3, 172–184. [CrossRef]

52. Ilyas, I.; Chu, X. Data Cleaning, 2nd ed.; Association for Computing Machinery: New York, NY, USA, 2019; p. 285. ISBN 978-1-4503-7152-0.

53. Jaziri-Bouagina, D.; Jamil, G. Handbook of Research on Tacit Knowledge Management for Organizational Success, 1st ed.; IGI Global: Hershey, PA, USA, 2017; p. 542. ISBN 978-1522523949.

54. Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; Manning, C.D. Stanza: A Python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the ACL: System Demonstrations, Online Conference, 5–10 July 2020; pp. 101–108. [CrossRef]

55. Musen, M.A. The protégé project: A look back and a look forward. AI Matters 2015, 1, 4–12. [CrossRef] [PubMed]

56. Bourque, P.; Fairley, R.E. Guide to the Software Engineering Body of Knowledge, Version 3.0; IEEE Computer Society: Washington, DC, USA, 2014; ISBN 978-0769551661.

57. Borges, P.; Monteiro, P.; Machado, R.-J. Mapping RUP Roles to Small Software Development Teams. In Software Quality. Process Automation in Software Development; SWQD 2012. Lecture Notes in BIP; Biffl, S., Winkler, D., Bergsmann, J., Eds.; Springer: Berlin, Germany, 2012; Volume 94. [CrossRef]

58. Kuhrmann, M.; Diebold, P.; Münch, J.; Tell, P.; Garousi, V.; Felderer, M.; Trektere, K.; McCaffery, F.; Linssen, O.; Hansen, E.; et al. Hybrid software and system development in practice: Waterfall, scrum, and beyond. In Proceedings of the 2017 ICSSP, Paris, France, 30–31 July 2017; pp. 30–39. [CrossRef]

59. Sfar, H.; Chaibi, A.H.; Bouzeghoub, A.; Ghelala, H.B. Gold standard based evaluation of ontology learning techniques. In Proceedings of the 31st Annual ACM Symposium on Applied Computing, Pisa, Italy, 4–8 April 2016; pp. 339–346. [CrossRef]

60. Raad, J.; Cruz, C. A Survey on Ontology Evaluation Methods. In Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Lisbon, Portugal, 12–14 November 2015; pp. 179–186. [CrossRef]

61. Cavalin, P.; Oliveira, L. Confusion matrix-based building of hierarchical classification. In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Proceedings of the Iberoamerican Congress on Pattern Recognition, Havana, Cuba, 28–31 October 2019; Springer: Cham, Switzerland, 2019; pp. 271–278. [CrossRef]

62. Ruuska, S.; Hämäläinen, W.; Kajava, S.; Mughal, M.; Matilainen, P.; Mononen, J. Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. Behav. Process. 2018, 148, 56–62. [CrossRef] [PubMed]