energies

MDPI

Article

SiSEG-Auto Semantic Annotation Service to Integrate Smart Energy Data

Maliheh Haghgoo *, Amirhossein Nazary Aghche Mazary and Antonello Monti

E.ON Energy Research Center, Institute for Automation of Complex Power Systems, RWTH Aachen University, Mathieustraße 10, 52074 Aachen, Germany; amirhossein.nazary@rwth-aachen.de (A.N.A.M.); amonti@eonerc.rwth-aachen.de (A.M.)

* Correspondence: mhaghgoo@eonerc.rwth-aachen.de

Abstract: In a modern smart energy system, the amount of available data from various sources is growing significantly. Other sectors such as medical or social sectors exhibit the same phenomenon. Due to the amount, complexity and heterogeneity of data, a complex algorithm is required for the integration and analysis of heterogeneous data sources. The Web of Things and semantic-based approaches address the fragmentation of standards, platforms, services and technologies in smart energy and non-energy sectors, and enable heterogeneous data integration and interoperability. This paper presents SiSEG, a semantic annotation service that is developed to automate the process of annotating data and address the problem of heterogeneous data integration in a reusable and extensible way by using the fuzzy method. Moreover, the accuracy of SiGEG has been evaluated.

Keywords: semantic annotation; smart energy data integration; web of things; heterogeneous data-set; semantic web; AI

Citation: Haghgoo, M.; Nazary Aghche Mazary, A.; Monti, A. SiSEG-Auto Semantic Annotation Service to Integrate Smart Energy Data. *Energies* **2022**, *15*, 1428. https://doi.org/10.3390/en15041428

Academic Editor: Gwanggil Jeon

Received: 17 November 2021
Accepted: 8 February 2022
Published: 15 February 2022

1. Introduction

The evolution in the smart energy system brings several challenges such as changes in energy demand, grid infrastructure, penetration of renewable energy sources, electric vehicles, and energy storage. From an Information and Communication Technology (ICT) perspective, the integration of the existing infrastructure in the smart energy domain requires efficient and cost-effective solutions. To ensure such systems are technically successful and widely adopted, ICT systems of different vendors and industrial companies must be able to integrate [1]. Additionally, these systems need to interact with platforms from various service providers for management and control [2]. Such a system is facing challenges and interoperability issues due to different representations of variant data sources.

A framework that has been used for interoperability purposes is the Smart Grid Architecture Model (SGAM), which is the main outcome of the Reference Architecture working group mandated by the EU's 490 Mandate [3]. Based on the SGAM framework there are five different layers of interoperability [4]: *Business layer* which represents the business view on the information (business models, market structures, business portfolios etc.). *Functional layer* specifies the functions and services. *Information layer* which is the data model and data semantics to be used to ensure a common understanding of the data exchanged. *Communication layer* which is the communication mechanism (e.g., PLC or Ethernet) and the communication protocol for data transmission. *Component layer* which is the physical distribution of all participating components to connect systems or devices. Base on SGAM, all interactions (i.e., physical, information-based and process-based) should satisfy the interoperability principle. This includes from the field level (e.g., substation automation, distribution automation and distributed energy resources), to remote operations (e.g., remote grid management), market management, service management, customer management and others.

The heterogeneity in communication, standardization and devices can be addressed by presenting comprehensive information of a domain [5]. According to the work in [5], a uniform view of the heterogeneous data sources has a direct impact on successful information sharing and searching of systems that are utilized for data integration. There are numerous applications in the smart energy and non-energy sectors that can benefit from integrated information. For example, integrated information of the measuring devices in the Smart Grid, traffic light sensors in the city, and user preferences can be used in a smart living system use-case to optimize the charging time of the electric vehicle. Furthermore, there exist other digitalized systems such as medical, agriculture, industry, education, etc. that can benefit from the integration of heterogeneous data sources in reporting and querying of existing services, for statistical analysis, online analytical processing, forecasting, visualizing, decision-making and planning.

The motivation for heterogeneous data sources integration is twofold [6]: to facilitate information access and reuse through a single information access point and to give certain information needed from a complementary information system. Furthermore, heterogeneity of data sources can be classified into four main categories [7]: structure, syntax, system and semantic.The structure heterogeneity refers to the different data models, syntax heterogeneity involves different languages and data representations. System heterogeneity is about hardware and operating system differences. The semantic heterogeneity is classified into three subcategories: *semantically equivalent concept* models use the same concept to present different terms, *semantically unrelated concepts* different system with completely different concepts uses the same term and *semantically related concept* generalization or specification of the same concept. A similar classification of heterogeneity can be found in [8].

The semantic web addresses the challenges of divers data integration for example for the smart energy system [9–11]. In particular, semantic web technologies have been employed to cross-cut domain-specific information and achieve a common understanding of information for humans besides providing machine-readable information.

In this interest, ontology [12] as a semantic web language is intended to provide rich and complex knowledge about things, groups of things, and relations between things. The term ontology has been used in many ways and domains such as medical, social, energy or education, [13–15]. In computer science, ontologies are introduced by Gruber [16] as an "explicit specification of a conceptualization". From this definition, an abstract model of how humans commonly think about real things in the word is called conceptualization. Furthermore, "explicit specification" describes the concepts and relations of the given abstract model with explicit name and definitions [17].

Based on the structure of an ontology, predicates are used to name and describe entities of a specific domain and their relationship. They represent the vocabulary at a conceptual level that can show knowledge about the domain and a set of the relationship. Therefore, ontologies and semantic web techniques are widely used to address data integration and solve heterogeneity problems [18].

Semantic technology enables blending data from disparate sources and further interlinking it as Resource Description Framework (RDF) statements into RDF triple-store that can be used in many knowledge management solutions. By adding semantics to the process of data integration, data pieces from disparate sources consolidate into meaningful and valuable data. In addition, being the backbone of semantic technology, RDF enables the inference of new facts from the existing data as well as the enrichment of the available knowledge by accessing Linked Open Data (LOD) resources [19].

The current smart energy system only provides end-to-end message delivery and lacks accessibility to semantic data. Organizations such as IETF, which manages Constrained Application Protocol (CoAP) standards, and Extensible Messaging and Presence Protocol (XMPP) are working on standardizing and integrating semantic data models into the protocols [20]. The semantic annotation technique can fill this gap of various

knowledge representations and the lack of semantic data representation. Relationships and interconnections can be established by applying semantic technologies.

Semantic annotation helps to add semantic meaning to the data and thereby describe it in a more concrete way.

In this paper, our vision is to develop an automated approach for the rapid development of interoperable smart energy applications using semantic web technologies, where rapid means faster application development with low-cost and less manual effort. To address the problem of heterogeneity of data sources, their integration and interoperability, an automated semantic annotation service called SiSEG is proposed. It can be plugged into any architecture and domain to provide interoperability and heterogeneous data integration between systems in an automated way using established communication and data standards. We present SiSEG as an ontology-based semantic annotation service to automatically add metadata to the raw data. Fuzzy analysis is used to automate the annotation process of the raw data.

This paper is organized as follows. An introduction to the smart energy system empowered by semantic technology, and a description of semantic annotation and fuzzy analysis are provided in Section 2. The implementation details of SiSEG are presented in Section 3. SiSEG is evaluated in Section 4 by using smart energy and IoT ontologies described in Section 2.1 as an exemplary use case. Section 5 concludes this paper.

2. Semantic as a Service in the Smart Energy System

The Semantic Annotation Service (SAS) is the heart of the semantic interoperability and heterogeneous data integration system that translates raw data and transfers it into a knowledge center application or platform as it is shown in Figure 1. Such a system facilitates interoperability and data integration at the data model level. In particular, to cross-cut domain-specific information and achieve a common understanding of information, semantic web technology has been engaged. At a high level, SAS connects external nodes via the support of MQTT, XMPP, or CoAP. On the other side, it interfaces with other cloud services via REST or a public protocol. The SAS acts as a mediator by processing raw data received from nodes and providing metadata at different levels according to the system requirements defined in an ontology.

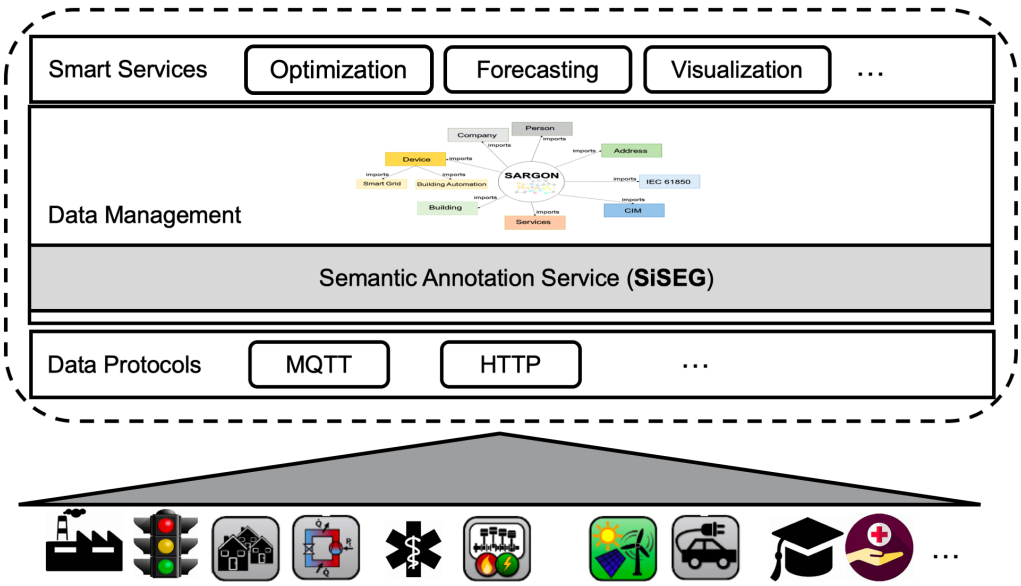

Figure 1. Semantic as a service in the smart system architecture.

2.1. Ontologies in Energy Domain

The revolution in the energy domain generates a large amount of data that is collected in real-time to give an extensive knowledge of the system. In such a system efficiency, interoperability and sustainability get more attention. Therefore, the smart energy system has been studied in recent research with a focus on changing the data representation and exchange technologies [21].

On the other hand, there have been several efforts for data exchange standardization in the smart home, energy and Internet of Things (IoTs) sectors. The European Telecommunications Standard Institute (ETSI) is currently working on a standardization activity and defines an open ontology called Smart Appliances REFerence (SAREF) [22]. SAREF enables information interoperability among IoT devices and servers using different technologies in several domains.

The most common domain-specific smart energy data use cases which have been explored in recent research projects using semantic ontologies are smart home, demand response management, energy managements and micro-grids. The most related ontologies have been summarized in [23–25]. In [24], the authors review existing ontologies in the smart energy domain and propose the SARGON ontology which is an extension of SAREF for cross-cutting domain-specific information and engaging smart grid with building energy automation sector. Figure 2 shows an overview of the covered network of ontologies that are interconnected via the core part of SARGON.

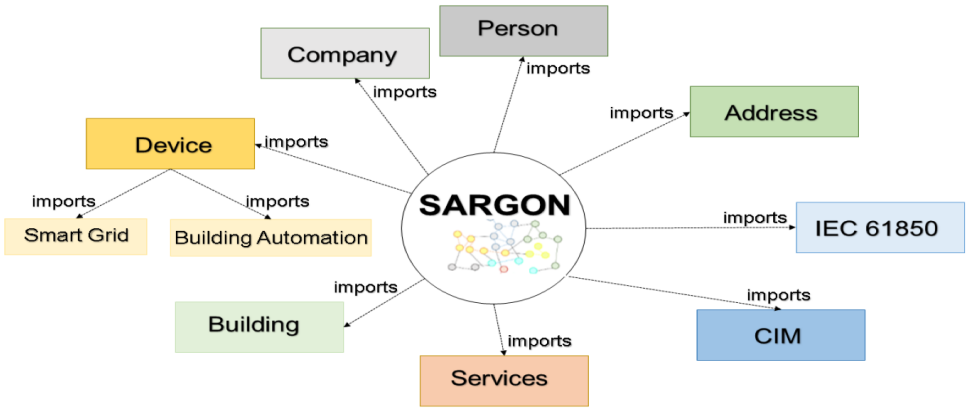

Figure 2. SARGON ontology network structure.

The modularity of SARGON enables its extension to any other standard and bridging information of different domains. SARGON divides the list of defined devices into two main categories of Building automation and Smart Grid. Two well-known standards IEC and Common Information Model (CIM) are linked into the core ontology.

In this study, SARGON has been selected as an example of a domain-specific ontology for annotating received data from the energy domain and stakeholders. The SAREF ontology is taken into account as a second domain-specific ontology and used for correctness validation in the assessment of the proposed tool.

2.2. Semantic Annotation Techniques

The provided raw data captured from heterogeneous data sources such as platforms, applications, services and devices do not contain any semantic and require a manual effort to build a more intelligent solution on top and provide information interoperability. However, due to the lack of annotation standards, even recent Internet of Things (IoT) services can not provide raw sensor data with included metadata [26].

Typically, smart energy applications are deployed in a bottom-up technique from sensor, over gateway and service to the application. Therefore, it depends on the data provider

how to control the sensor data and data structure. Consequently, domain information has turned into various vertical sub-domains with no horizontal connectivity between them.

The lack of interoperability in such a system is a clear disadvantage for smart energy applications that can benefit from multiple devices and data providers. To address the interoperability issue while integrating heterogeneous data of various vertical domains, raw data can be utilized by semantic annotation which uses standard mechanisms and vocabulary between different data providers to normalize the data.

Annotating data is one of the major techniques to create metadata and put machine-understandable data on the web. Moreover, the annotation of data is a costly and time-consuming task. Therefore, using efficient approaches for the annotation tool is significant for the performance.

The tools for annotating data can be categorized into three main groups [27]: manual semantic annotations, semi-automatic annotation and automatic annotation. Many different functionalities are required to make the annotation process as automatic as possible.

The data annotation schemes described in [28] transfer raw data to RDF based on a provided scheme.

Besides the schemes description, the source mapping and annotation process have been presented. However, such an approach does not always fit for only one specific schema and is incompatible with heterogeneous data sources and vendors. Therefore, there have been several efforts to address the task of semi-automatic and automatic data annotation and query. In the following, some of recent and related techniques have been described.

Flexible and novel data integration has been introduced in [29] to annotate heterogeneous data using domain ontologies. However, semantic relations among heterogeneous data are not effectively illustrated in this method. The work in [30] classifies the information into no domain-specific and domain-specific with using linked data. This technology is limited to the medical relevant databases. In [31], a weather monitoring system that stores the sensor data in real-time and handles spatial and temporal queries is presented. However, this research is also focused on querying historical data. The author of [32] employed linked data to improve telecommunication operation by enriching the textual documents. ReDy Artificial Intelligent (AI) method is used to engage mobile service with the web of data. Therefore, the semantic annotation process utilizes the advantage of information that is presented on the web. Nevertheless, multi-domains and cross-cutting domain-specific information require linked data concepts. The crowdsourced semantic annotation (SemAnn) [33] introduces a common collaborative annotation of text and tables by using the hierarchical context of annotations, although it is mainly used for the text annotation and limited to the DBpedia resources.

To cope with the aforementioned issues and improve scalability, the work in [34] integrated the web of things (WoT) concept to include both manual and semi-automatic annotators. The introduced method can be improved with unsupervised learning instead of using supervised learning algorithms.

To tackle the problem of scalability, Semantic Annotation over Summarized sensOr Data stReam (SEASOR) [35] has been introduced. It uses a sub-window partitioning method and extended semantic sensor network (SSN). However, SEASOR is limited to the SSN domain.

There exists a limited list of automated annotation tools support for multiple vocabularies. There is a general lack of freely available simple annotation tools that are not of specialized use and limit the user to a specific ontology. Therefore, an automated and modular solution address the existing gaps which inspire the development of SiSEG.

2.3. Fuzzy Method

The automated learning of models from experimental data is a core aspect of machine learning. In machine learning, a set of data is used to produce a learning algorithm. It takes a data set $z \in Z$ where Z is Cartesian of a fixed set of attributes. The observation z can be described in terms of a feature vector and is aimed to find any possible structure in the data. In contrast to the traditional knowledge-driven approach, in machine learning,

several complementary data-driven methods can be applied. Among those, fuzzy analysis has been introduced as a data-driven adaptation method [36].

Fuzzy analysis is a method for solving problems which are related to uncertainty and vagueness. It is used in multiple areas, such as engineering and has applications in decision making problems, planning and production [36]. Over the past years, the fuzzy analysis method has attracted attention for automated learning and the extraction of patterns from experimental data [37]. In addition to this, it has been focused on the knowledge discovery in database (KDD) as a response to the progress of digital data acquisition where there exist limited human capabilities to analyze and explore large amounts of data.

Fuzzy analysis derives from the mathematical study of multi-valued logic. In a normal logic, only the absolute truth is taken into account. Therefore, fuzzy analysis helps to mimic the way humans analyze problems and make decisions, in a way that relies on vague or imprecise values rather than absolute truth or falsehood [37].

SiSEG applies fuzzy analysis for ranking of extracted words and Uniform Resource Identifiers (URIs). Decisions rely on vague and imprecise values rather than absolute true and false. The annotation process in SiSEG is an uncertain decision as it integrates diverse data that might not match with its knowledge base. SiSEG tries to find the closest information by using fuzzy analysis.

3. SiSEG Implementation

All of the already existing annotation tools and services are semi-automatic or limited to a specific ontology. To integrate heterogeneous data sources and provide interoperability, an automated annotation service is beneficial to normalize the represented data without limitation to the specific ontologies. Therefore, SiSEG implements an automated procedure for annotating data. It receives heterogeneous raw data from the energy sector and also non-energy sectors and adds metadata to raw data to generate annotated and normalized data according to its knowledge base. SiSEG's knowledge base is not limited to the energy sector and can include other sectors as well. The fuzzy method is used in SiSEG for automating decisions and generating annotated data.

An overview of the semantic annotation service is shown in Figure 3. As inputs, SiSEG gets raw data in JSON or XML format and required ontologies as domain knowledge. After processing and tagging the data, it generates annotated data in JSON-LD, RDF and OWL formats. This annotated data can be used in any use cases where heterogeneous data integration is required.

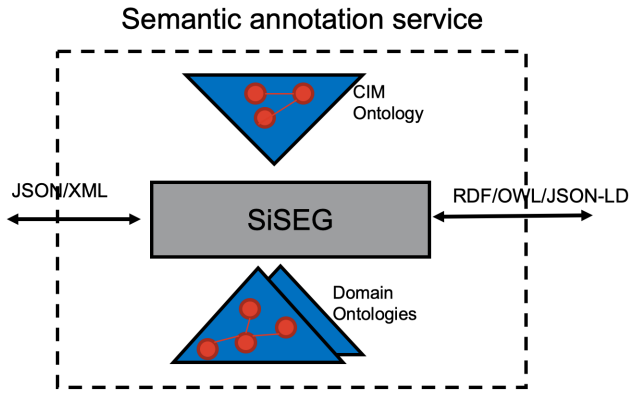

Figure 3. Semantic annotation service overview.

According to Section 2.1, in this study, SAREF for smart appliances and SARGON for the smart energy domain are employed as knowledge base ontologies. Nevertheless, SiSEG is valid for other knowledge base ontologies and ontologies form other sectors can be considered.

The data annotation block which shows the main functionality of SiSEG as a core part of the procedure is detailed in Figure 4. The procedure of automating annotation of raw data has four steps:

1. Keyword Extraction
2. URI Extraction
3. Feature Vector Generation
4. and Support Vector Machine (SVM)

Following this four-step process, a semantically annotated file in JSON-LD, RDF and OWL format are generated as output.

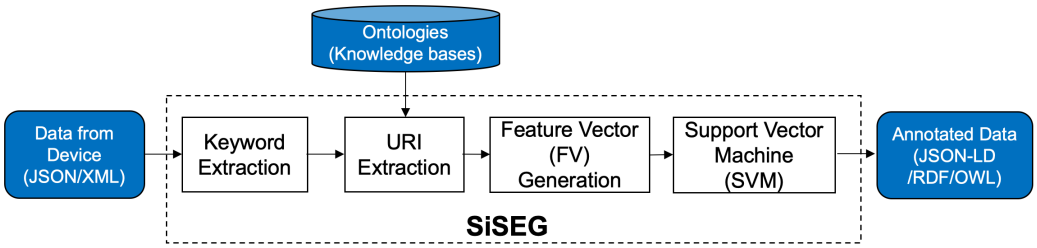

Figure 4. SiSEG functions and process overview.

3.1. Step 1: Keyword Extraction

In order to annotate inputs with the most relevant URIs, it is important to extract all keywords from the input [38]. Thus, in this step, keywords are extracted from raw data for further annotation purposes. In case the raw data has a specific data-interchange format (for example JSON), keywords will include all the strings in the JSON file. As JSON is a format consisting of key and value pairs in which keys are strings and values will lead to strings as well, this step will be a straightforward process of extracting all the strings in a JSON file, splitting them by space sign and saving the split strings in a list as keywords. The same process can be carried out if the raw data has XML format. Other helpful pieces of data, such as data types of values in a JSON file, are extracted in this step. They will come into use in further steps where we intend to filter irrelevant results or correct the input. In case the raw data does not have any specific format or is a long string containing data to be transferred, splitting the long string by space sign into small words is how keywords are extracted. Raw data lacking a data-interchange format may cause lower accuracy of the whole annotation process, thus a condition in which raw data have a specific format (like XML or JSON) is considered advantageous comparing to raw data having no specific format at all.

3.2. Step 2: URI Extraction

The second step includes querying data from knowledge bases. For this purpose, the SPARQL query language is used to query over knowledge base models. The output of the SPARQL query is a result set containing URIs or RDF nodes that might be relevant to the keywords which were extracted in the previous step. A simple example of SPARQL query over SAREF ontology is depicted in Figure 5. The query request selects subject nodes which are a subclass of `saref:Device` node in the ontology. The result of the query request is illustrated in Figure 6.

All the keywords which were extracted in step 1 are searched in the knowledge base to find URIs or RDF nodes which are related to the extracted word. However, only querying exact keywords from the knowledge base might not show all the desired URIs. As an example, consider a case in which *TemperatureSensor* is extracted as a keyword in step 1. Only searching the exact word will show limited URIs which are directly explaining "temperature sensor" and not showing other URIs which might be relevant, such as URIs about "temperature" and "sensor" separately These are not the exact keyword, but might be relevant none the less. As a solution to this issue, all morphemes of the keyword

are extracted based on morphemes tree method [39] to analysis the word structure and searched over the knowledge base. This way, words such as "temperature" and "sensor" will also be queried from the knowledge base and, thus, relevant URIs will be included in the result set. Figure 7 illustrates this process. Obtained results will be further analyzed and filtered in the next steps.

```
PREFIX saref: <https://saref.etsi.org/core/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?subject
WHERE {
   ?subject rdfs:subClassOf saref:Device .
} Limit 5
```

Figure 5. SPARQL query over SAREF to find `Devices` in the ontology.

```
saref:Sensor
saref:Meter
saref:Appliance
saref:Actuator
saref:HVAC
saref:TemperatureSensor
```

Figure 6. Result for SPARQL query.

3.3. Step 3: Feature Vector Generation

In step 3, queried data (URIs or RDF Nodes) available from the previous step are prepared for further analysis by assigning vectors to them which are called Feature Vectors. Each Feature Vector has two elements comprised of two float values which are obtained through the following methods based on fuzzy analysis:

- Surface Similarity
- Popularity

Surface Similarity value is calculated by measuring string similarity between input and the label or comments assigned to the URI or RDF subject node in the ontology. By comparing input and the mentioned labels or comments, a float value ranging from 0 to 1 is attained which is called Surface Similarity value. This value will then be placed as the first element of the Feature Vector assigned to the URI.

The second element of Feature Vector includes the Popularity value of the URI. The popularity value is calculated by obtaining the repetition number of the URI in the Query result set of step 2. The more a URI is repeated in the result set, the higher its popularity value will be.

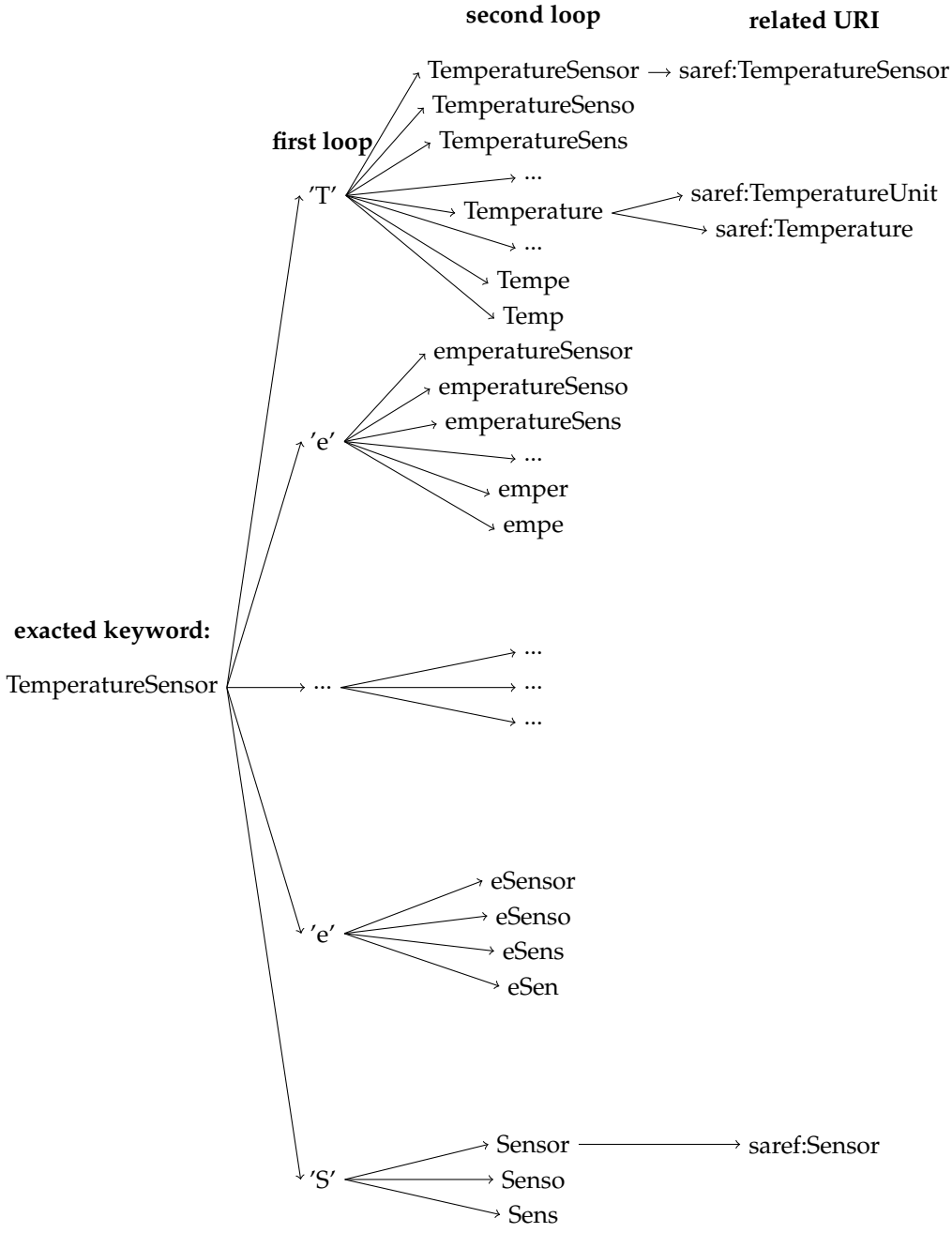

Figure 7. Extracting morphemes of a keyword.

Based on the dynamic and modular design of SiSEG, Feature Vectors can be modified to achieve higher precision with minimum change to the rest of the tool. For this purpose, alongside Surface Similarity and Popularity, new methods (represented by a float value) can be appended to the tool. For this purpose, one of the techniques that can be employed is Word2vec. Word2vec provides a numerical representation of each distinct word in the form of a vector [40]. By comparing Word2vec vectors of keywords in the input with the labels of URIs available in knowledge bases [41], a single float number is obtained which can be appended into the already defined Feature Vector. In such a case, Feature Vectors will be modified into higher dimensions while other steps require no change.

3.4. Step 4: Support Vector Machine

The fourth step involves analyzing and filtering URIs. A Support Vector Machine (SVM) is used in order to find URIs which are related the most to the input. This task is done by the SVM using fuzzy method to classify each URI as relevant or irrelevant to the input. Feature Vectors corresponding to each URI are used for this purpose. Initially, the SVM has to be trained with a set of training data. Subsequently, a decision line is obtained which is the main factor for further classification decisions. A simple example for the SVM's training process and the decision line as the output is illustrated in Figure 8.

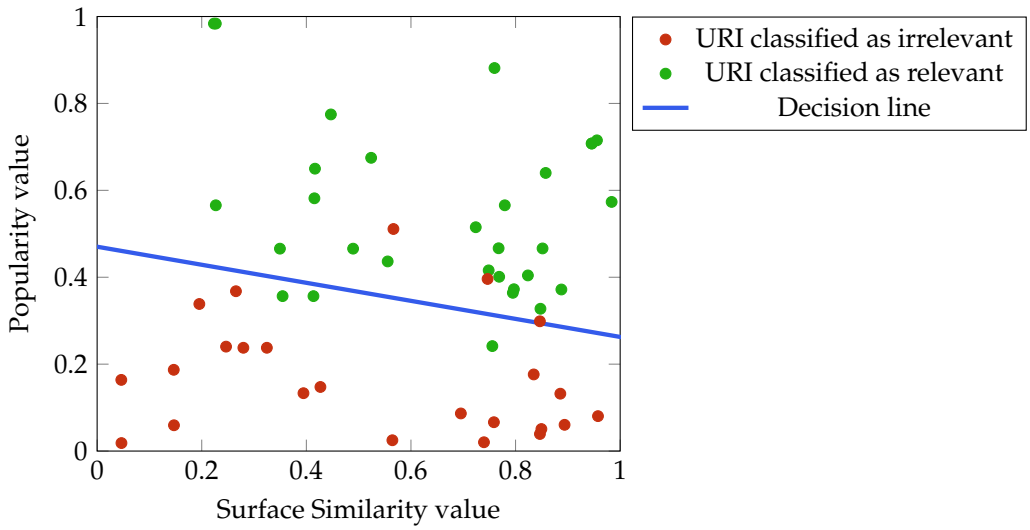

Figure 8. SVM's training data and corresponding decision line.

After the SVM training process ends, it can be used for classifying the first step's result as relevant or irrelevant to the input. Only URIs which are classified as relevant will be used in further steps and the rest will be tagged as irrelevant data and will be excluded from the output.

As the training process for SVM ends, URIs are sent to SVM as inputs and afterward, they are classified as either relevant or irrelevant. URIs that are classified as relevant will then be compared to each other considering their assigned Feature Vectors. Among a set of relevant-classified URIs, ones with the highest distance to SVM's decision line are chosen as the most relevant data to input, thus they are included in the output JSON-LD file.

4. SiSEG Assessments

The analysis in this Section aims at evaluating the accuracy of SiSEG in terms of the annotation and harmonization of raw data. Therefore, a list of random devices has been selected to conduct the assessment of SiSEG. In this assessment, all tests were performed offline on twenty-two random types of devices that are not limited to the energy sector to address heterogeneous data sources integration. Further information regarding the type of devices is given in Appendix A.

In addition, SiSEG performs not only the data publishing but also the context information process to handle annotation steps that are already described in Section 3. However, in this assessment, the accuracy of produced data has experimented and the computation time of the process was not focused on. In the following Section, detail of assessment method has been described.

4.1. SiSEG Assessment Method

To assess the produced annotated data, the result of SiSEG which is a generated JSON-LD file for each device has been analyzed. To consider integration of heterogeneous data, twenty-two dummy and randomly generated raw data sources with the same level

of complexity corresponding to the different types of devices are sent to SiSEG as input. Complexity can be identified according to the number of properties, relationships, meta information and length of a message. In this process, all raw data were annotated using both SAREF and SARGON ontologies as knowledge bases separately. At the end of the process, SiSEG saves an annotated file that refers to a certain device with related properties. Furthermore, to assess the results of SiSEG and identify its accuracy, the output of SiSEG is categorized into three groups of relevant URIs, partially relevant URIs and irrelevant URIs. The following description is given for each category:

Relevant URI An output file is categorized here if all URIs in the file match with a device description from the knowledge base.

Irrelevant relevant URI An output file is categorized here if all URIs in the file do not match with a device description from the knowledge base.

Partially URI An output file is categorized here if URIs in the file partly match with a device description from the knowledge base.

Furthermore, as it is applied in Algorithm 1, the accuracy has been calculated by dividing total number of relevant URIs to sum of relevant URIs, partially relevant URIs and irrelevant URIs.

Algorithm 1 Calculating accuracy.

Require: $URIs \geq 0$
Require: $X, Y, Z \geq 0$
  while $URIs \neq 0$ **do**
    if $URIs$ is relevant **then**
      $X \leftarrow X + 1$
    else if $URIs$ is not relevant **then**
      $Y \leftarrow Y + 1$
    else if $URIs$ is partially relevant **then**
      $Z \leftarrow Z + 1$
    end if
  end while
  $Accuracy \leftarrow \frac{X}{X+Y+Z}$

4.2. SiSEG Assessment Result

Table 1 presents the average number of relevant URIs as a measure for the accuracy of SiSEG in annotation of variant devices. The computation of the accuracy is done with Algorithm 1.

Table 1. SiSEG assessment result.

Different Test Cases	Relevant URIs (%)
No specifications in knowledge base	82.39%
Device defined in knowledge base	96.77%
Device not defined in knowledge base	27.27%
Using only SAREF as knowledge base	63.63%
Using only SARGON as knowledge base	95.45%

SiSEG reaches an overall accuracy of 82.39% in finding relevant URIs using SAREF and SARGON together as knowledge base. As the list of devices was selected randomly and not completely defined in the knowledge bases, this result was expected. Furthermore, the average accuracy of finding relevant URIs is 95.45% if only SARGON is used as knowledge base and 63.63% for SAREF. Such a difference can be seen between SAREF

and SARGON, due to the fact that SARGON ontology is more general than SAREF and including diverse devices.

In cases where the raw data describe a device that is defined in the knowledge base, the average accuracy increases to 96.77%, while in cases where the raw data describes a device that is not explicitly defined in the knowledge base, the average overall accuracy decreases to 27.27%. As was predicted, this gap becomes smaller if devices or tested data sets are known in the knowledge base. SiSEG is proposed for heterogeneous data integration and it is modular. This accuracy can be varied between 27.27% to 96.77% depending on the data set and knowledge base. Of course, accuracy can stay at 96.77% if the data set will be known to the knowledge base.

According to the assessment result, it is shown that, if a list device that is known or partly known in the knowledge base will be evaluated, then the accuracy of SiSEG will stay similar to the given assessment. Moreover, the accuracy will drop if the tested devices are not defined in the knowledge base. The accuracy of SiSEG depends on the type of devices and the given knowledge bases. However, both limitations are acceptable since SiSEG is modular and reusable. Therefore, the knowledge base of SiSEG can be modified to include a more relevant list of devices to enhance the accuracy.

Moreover, the result shows that SiSEG is more accurate than the comparable annotation tool SemAnn (see Section 2.2 and [34]). According to the article [34], the accuracy of SemAnn is 40% for closely related, 30% for vaguely related, and 30% for unrelated. The accuracy of SemAnn can increase to 97% depending on the number of recalling URIs. SiSEG achieves the better accuracy than SemAnn as it does not depend on the number of recalling URIs from knowledge base to annotate the data.

5. Conclusions

Heterogeneous data integration in smart energy and non-energy systems requires intelligent and interoperable solutions to conduct a uniform and harmonize data representation. In this study, semantic technologies are employed to harmonize the received raw data from diverse resources and facilitate accessing and integrating heterogeneous data. To foster the harmonization process, an automated service for semantic annotation is required. However, there exists a limited list of automated annotation services and tools that support for multiple domain ontologies. Therefore, in this article, SiSEG as a modular service has been proposed and its development procedure has been described to address the existing gaps in automating the process of semantic annotation. SiSEG is a service and solution for integrating energy and non-energy sectors where heterogeneous data integration is beneficial.

Furthermore, the accuracy of SiSEG is evaluated in terms of finding relevant and meaning in the annotated result. Based on the result, SiSEG is more accurate than SemAnn which is a comparable tool for semantic annotation with SiSEG. Although, the accuracy of SiSEG depends on the type of devices and taken knowledge base. These dependencies are acceptable as SiSEG is a modular service and user can adopt different knowledge bases to achieve higher accuracy. Moreover, observed accuracy shows that if we use SARGON and SAREF as knowledge bases, it is promising to bring SiSEG in real and practical use-cases in several sectors such as medical, education, industrial, agriculture, etc.

The modularity of SiSEG provides a wide operational domain as its knowledge base can be changed according to the required domain. Furthermore, SiSEG can be extended to provide more features to the service. Besides the improvement suggestions from the evaluation, we consider the most important future work to be in extending SiSEG to store and publish the annotated data in the Open Annotation ontology server [42] and optimize the defined method to increase the accuracy of SiSEG even when using unknown devices and more complex scenarios.

Author Contributions: Conceptualization, M.H.; Formal analysis, M.H.; Methodology, M.H.; Software, M.H. and A.N.A.M.; Supervision, A.M.; Writing—original draft, M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work is part of project "OneNet", this project has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under grant agreement no 957739.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

IoTs	Internet of Things
WoT	Web of Things
SGAM	Smart Grid Architecture Model
RDF	Resource Description Framework
LOD	Linked Open Data
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
CoAP	Constrained Application Protocol
XMPP	Extensible Messaging and Presence Protocol
MQTT	Message Queuing Telemetry Transport
ETSI	European Telecommunications Standard Institute
SAREF	Smart Appliances REFerence Ontology
SARGON	Smart Energy System Ontology
CIM	Common Information Model
SAS	Semantic Annotation Service
AI	Artificial Intelligence
SEASOR	SEmantic Annotation over Summarized sensOr Data stReam
SSN	Semantic Sensor Network ontology
KDD	Knowledge Discovery in Database
SiSEG	Automatic SemantIc annotation Service for smart Energy data inteGration
FV	Feature Vector
URI	Uniform Resource Identifier
SVM	Support Vector Machine

Appendix A

The list of tested devices is Actuator, Floor, RVK, Light Switch, Temperature, Meter, Smoke, Door Switch, Current Meter, Wash Machine, Water Flow Meter, PV, HVAC, Room, Lighting Device, Building, Charging Station, Traffic, Pollution, Weather, Battery Storage, PMU. Furthermore, Figures A1 and A2 present an example of received raw data and annotated data for a Floor.

```
{  "id": "Floor:01",
    "type": "Floor",
    "inBuilding": {
        "type": "Relationship",
        "object": "Building:01"},
   "hasRoom": {
         "type": "Relationship",
        "object": [
            "Room:01",
            "Room:02"]},
   "hasZone": {
         "type": "Relationship",
        "object": [
            "Zone:01",
            "Zone:02"]}
},
  "@context": [
         "https://.../Context-Information/Core-Context.jsonld",
         "https://.../Context-Information/Building.jsonld"
  ]
}
```


Figure A1. Annotated data of Floor.


```
{
"Floor:01",
"inBuilding":"Building:01",
"hasRoom":"Room:01,02",
"hasZone":"Zone:01, 02"
}
```


Figure A2. Raw data of Floor.

References

1. Da Xu, L.; He, W.; Li, S. Internet of things in industries: A survey. *IEEE Trans. Ind. Inf.* **2014**, *10*, 2233–2243.
2. Ray, P.P. A survey of iot cloud platforms. *Future Comput. Inform. J.* **2016**, *1*, 35–46. [CrossRef]
3. Smart Grid Mandate: Standardization Mandate to European Standardisation Organisations (ESOs) to Support European Smart Grid Deployment, DG ENER, European Commission. March 2011. Available online: https://ec.europa.eu/energy/sites/ener/files/documents/2011_03_01_mandate_m490_en.pdf (accessed on 25 August 2021).
4. CEN-CENELEC-ETSI Smart Grid Coordination Group "Smart Grid Reference Architecture". 2012. Available online: https://ec.europa.eu/energy/sites/ener/files/documents/xpert_group1_reference_architecture.pdf (accessed on 25 September 2021).
5. Bandyopadhyay, D.; Sen, J. Internet of things: Applications and challenges in technology and standardization. *Wirel. Pers. Commun.* **2011**, *58*, 49–69. [CrossRef]
6. Ziegler, P.; Dittrich, K.R. Data integration—Problems, approaches, and perspectives. In *Conceptual Modelling in Information Systems Engineering*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 39–58.
7. Kim, W.; Seo, J. Classifying schematic and data heterogeneity in multidatabase systems. *IEEE Comput.* **1991**, *24*, 12–18. [CrossRef]
8. Goh, C.H. Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Sources. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1997. Available online: http://ccs.mit.edu/ebb/peo/mad.html-03/2003 (accessed on 25 September 2021).
9. Curry, E.; Hasan, S.; O'Riain, S. Enterprise energy management using a linked dataspace for energy intelligence. In Proceedings of the 2012 Sustainable Internet and ICT for Sustainability (SustainIT), Pisa, Italy, 4–5 October 2012; pp. 1–6.
10. Kofler, M.J.; Reinisch, C.; Kastner, W. A semantic representation of energy-related information in future smart homes. *Energy Build.* **2012**, *47*, 169–179. [CrossRef]
11. Daniele, L.; Solanki, M.; Hartog, F.D.; Roes, J. Interoperability for smart appliances in the IoT world. In Proceedings of the International Semantic Web Conference, Kobe, Japan, 17 October 2016; pp. 21–29.
12. Gómez-Pérez, A.; Corcho, O. Ontology languages for the semantic web. *IEEE Intell. Syst.* **2002**, *17*, 54–56. [CrossRef]

13. 3rd Millennium, Inc. Practical Data Integration in Biopharmaceutical R&D: Strategies and Technologies. A White Paper. May 2002. Available online: http://www.3rdmill.com/ (accessed on 25 September 2021).
14. Chandrasekaran, B.; Josephson, R. What are ontologies, and why do we need them? *IEEE Intell. Syst. Their Appl.* **1999**, *14*, 20–26. [CrossRef]
15. Event/Process-Based Data Integration for the Gulf of Maine. Campobello Island, New Brunswick. 12–14 June 2002. Available online: www.spatial.maine.edu/bdei/bdeippr.pdf-03/2003 (accessed on 25 September 2021).
16. Gruber, T. A translation approach to portable ontology specifications. *Knowl. Acquis.* **1993**, *5*, 199–220. Available online: http://ksl-web.stanford.edu/KSL_Abstracts/KSL-92-71.html-07/2003 (accessed on 25 May 2021). [CrossRef]
17. Visser, U.; Schlieder, C. Modelling with Ontologies. In *The Ontology and Modeling of Real Estate Transactions*; Routledge: London, UK, 2002. [CrossRef]
18. Cui, Z.; O'Brien, P. Domain Ontology Management Environment. In Proceedings of the 33rd Hawaii Annual Hawaii International Conference on System Sciences, Maui, HI, USA, 7 January 2000.
19. Halevy, A.; Rajaraman, A.; Ordille, J. Data integration: The teenage years. In Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, 12 September, 2006; pp. 9–16.
20. Waher, P. Sensor Data Interchange over XMPP. Available online: https://xmpp.org/extensions/inbox/sensors.html (accessed on 25 May 2021).
21. D'Onofrio, S.; Portmann, E. Cognitive computing in smart cities. *Inform.-Spektrum* **2017**, *40*, 46–57. [CrossRef]
22. Daniele, L.; den Hartog, F.; Roes, J. Created in Close Interaction with the Industry: The Smart Appliances REFerence (SAREF) Ontology. In *International Workshop Formal Ontologies Meet Industries*; Springer: Cham, Switzerland, 2015; Volume 225, pp. 100–112.
23. Wierling, A.; Schwanitz, V.J.; Altinci, S.; Bałazińska, M.; Barber, M.J.; Biresselioglu, M.E.; Burger-Scheidlin, C.; Celino, M.; Demir, M.H.; Dennis, R.; et al. Fair metadata standards for low carbon energy research—A review of practices and how to advance. *Energies* **2021**, *14*, 6692. [CrossRef]
24. Haghgoo, M.; Sychev, I.; Monti, A.; Fitzek, F.H. SARGON–Smart energy domain ontology. *IET Smart Cities* **2020**, *2*, 191–198. [CrossRef]
25. Cavalieri, S. Semantic Interoperability between IEC 61850 and oneM2M for IoT-Enabled Smart Grids. *Sensors* **2021**, *21*, 2571. [CrossRef] [PubMed]
26. Noura, M.; Atiquzzaman, M.; Gaedke, M. Interoperability in internet of things: Taxonomies and open challenges. *Mob. Netw. Appl.* **2019**, *24*, 796–809. [CrossRef]
27. Erdmann, M.; Maedche, A.; Schnurr, H.P.; Staab, S. From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. In Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content, Luxembourg, 5–6 August 2000.
28. Zhang, X.; Zhao, Y.; Liu, W. A method for mapping sensor data to SSN ontology. *Int. J. u-e-Serv. Sci. Technol.* **2015**, *8*, 303–316. [CrossRef]
29. Boury-Brisset, A.C. Managing Semantic Big Data for Intelligence. In Proceedings of the 8th International Conference on Semantic Technologies for Intelligence, Defense, and Security, Fairfax, VA, USA, 12–15 November 2013.
30. Zhao, J.; Sahoo, S.S.; Missier, P.; Sheth, A.; Goble, C. Extending semantic provenance into the web of data. *IEEE Internet Comput.* **2010**, *15*, 40–48. [CrossRef]
31. Ha, S.W.; Lee, Y.K.; Vu, T.H.N.; Jung, Y.J.; Ryu, K.H. An environmental monitoring system for managing spatiotemporal sensor data over sensor networks. *Sensors* **2012**, *12*, 3997–4015. [CrossRef]
32. Rocha, O.R.; Vagliano, I.; Figueroa, C.; Cairo, F.; Futia, G.; Licciardi, C.A.; Marengo, M.; Morando, F. Semantic annotation and classification in practice. *IT Prof.* **2015**, *17*, 33–39. [CrossRef]
33. Takis, J.; Islam, A.S.; Lange, C.; Auer, S. Crowdsourced semantic annotation of scientific publications and tabular data in PDF. In Proceedings of the 11th International Conference on Semantic Systems, Vienna, Austria, 16–17 September 2015.
34. Wu, Z.; Xu, Y.; Zhang, C.; Yang, Y.; Ji, Y. Towards Semantic web of things: From manual to semi-automatic semantic annotation on web of things. In *International Conference on Big Data Computing and Communications*; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 295–308.
35. Pacha, S.; Murugan, S.R.; Sethukarasi, R. Semantic annotation of summarized sensor data stream for effective query processing. *J. Supercomput.* **2020**, *76*, 4017–4039. [CrossRef]
36. Coroiu, A.M. Fuzzy methods in decision making process—A particular approach in manufacturing systems. *IOP Conf. Ser. Mater. Sci. Eng.* **2015**, *95*, 012154. [CrossRef]
37. Hüllermeier, E. Fuzzy methods in machine learning and data mining: Status and prospects. *Fuzzy Sets Syst.* **2005**, *156*, 387–406. [CrossRef]
38. Gonnerman, L.M. A linguistic analysis of word morphology. In *Morphological Processing and Literacy Development*; Routledge: New York, NY, USA, 2018; pp. 3–15.
39. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 1–3.
40. Nadim, I.; El Ghayam, Y.; Sadiq, A. Semantic Annotation of Web of Things Using Entity Linking. *Int. J. Bus. Anal.* **2020**, *7*, 6–10. [CrossRef]

41. Marques, O.; Barman, N. Semi-automatic Semantic Annotation of Images Using Machine Learning Techniques. In Proceedings of the Semantic Web-ISWC 2003, Second International Semantic Web Conference, Sanibel Island, FL, USA, 20–23 October 2003; pp. 554–558.
42. Ciccarese, P.; Clark, T. Annotopia: An Open Source Univer- sal Annotation Server for Biomedical Research. In Proceedings of the 7th International Workshop on Semantic Web Applications and Tools for Life SWAT4LS, Berlin, Germany, 9–11 December 2014.