CLIMA 2022
Eye on 2030 | Towards digitalized, healthy, circular and energy efficient HVAC

REHVA 14th HVAC World Congress
22nd – 25th May, Rotterdam, The Netherlands

Interoperability of semantically heterogeneous digital twins through Natural Language Processing methods

Alina Cartus [a], Maximilian Both [a*], Nicolai Maisch [a], Jochen Müller [a], Christian Diedrich [b].*

*First Authors

[a] Institute of Building Services Engineering, TH Köln, University of Applied Sciences, Cologne, Germany, {acartus; maximilian_alexander.both; nicolai.maisch; jochen.mueller}@th-koeln.de.

[b] Institute of Automation Technology, Otto von Guericke University Magdeburg, Magdeburg, Germany, christian.diedrich@ovgu.de.

Abstract. Self-organizing systems represent the next stage in the development of automation technology. For being able to interact with each other in an interoperable manner, it requires a uniform digital representation of the system's components, in the form of digital twins. In addition, the digital twins must be semantically interoperable in order to realize interoperability without the need for costly engineering in advance. For this purpose, the current research approach focuses on a semantically homogeneous language space. Due to the multitude of actors within an automation network, the agreement on a single semantic standard seems unlikely. Different standards and vendor-specific descriptions of asset information will continue to exist. This paper presents a method extending the homogeneous semantics approach to heterogeneous semantics. For this purpose, a translation mechanism is designed. The mapping of unknown vocabularies to a target vocabulary enables the interactions of semantically heterogeneous digital twins. The mapping is based on methods from the artifcial intelligence domain, specifically machine learning and natural language processing. Semantic attributes (name, definition) as well as further classifying attributes (unit, data type, qualifier, category, submodel element subtype) of the digital twins' attributes are used therefore. For the mapping of the semantic attributes pre-trained language models on domain specific texts and sentence embeddings are combined. A decision tree classifies the other attributes. Different semantics for submodels of pumps and HVAC systems are used as the evaluation dataset. The combination of the classification of the attributes (decision tree) and the subsequent semantic matching (language model), leads to a significant increase in accuracy compared to previous studies.

Keywords. Semantic interoperability, Natural Language Processing, Decision Tree, Asset Administration Shell
DOI: https://doi.org/10.34641/clima.2022.143

1. Introduction

The change to self-organizing systems is shaping the next development stage of automation technology. In automation technology, the digital, global networking of systems with self-x capabilities is being pushed. Self-x capabilities are understood as functionalities of a system that enable intrinsic automatisms for network exploration, self-configuration, -diagnosis and -optimization. Self-x capabilities of systems enable interoperability and automation based on it. Interoperability means that systems actively collaborate across specific product, system, and process boundaries to meet common functional fulfillments [1, 2]. If this cooperation is automated, manual configuration and control efforts are reduced and result in an optimization of the value chain.

If interoperability is guaranteed for systems of technical building equipment (TBE), a self-configuration of interactions (e.g. the integration of energy values into a monitoring application) can be realized by means of automatically executable rules and engineering efforts can be avoided or essentially reduced. A prerequisite for the interoperability of systems is their digital, uniform representation, as well as the guarantee of semantic interoperability (SI) of the systems. SI refers to the ability to correctly exchange data among themselves and to understand what they mean [3].

To achieve SI, the current research approach focuses on semantically homogeneous descriptions of systems [2, 4]. This homogeneity comes from standardization and harmonization activities of information for different components of industry and TBE (e.g. [5–8]). If, for example, properties for the power consumption of different components are based on a uniform vocabulary, they can be automatically integrated into a monitoring application (Figure 1) [9].

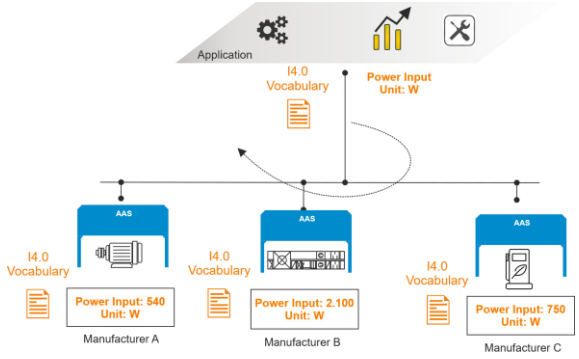

Fig. 1 – Semantically homogeneous AAS

However, if systems are based on heterogeneous vocabularies, interactions must be configured through manual effort and expert knowledge. This is the current state of the art for the majority of industrial and building applications. Because of the effort involved, operators carefully weigh the implementation of multi-vendor applications, such as plant asset management applications, against their benefits. The high configuration effort stands in the way of the wide availability of these applications. [10]

This paper contributes to extending the research approach from interactions of semantically homogeneous to heterogeneous systems. The interoperability of semantically heterogeneous systems can be achieved by mapping heterogeneous descriptions, to the, underlying semantic standard of the respective components. This paper presents the method of automated matching, which allows systems to map semantically heterogeneous descriptions to their own standard independently and without configuration (Figure 2).

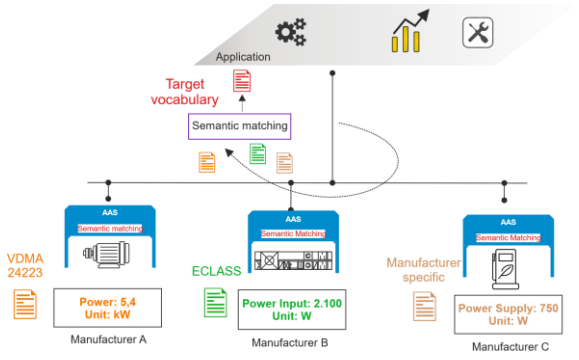

Fig. 2 – Semantically heterogeneous AAS

The starting point for matching is the information from the digital representations of the components. These are available in a uniform structure, but do not follow a semantically uniform standard. Methods from Natural Language Processing (NLP) [11] serve as the basis for matching semantic attributes.

In a first draft [12] the semantic attributes "name" and "definition" of heterogeneously labeled pump properties from the project [8] were used for automated matching. The mapping of heterogeneous semantics to a defined target vocabulary was achieved using a pretrained language model (PLM). For technical language understanding, extended pretraining of the PLM was performed with technical literature (Step 1, Figure 3). Combined with sentence embeddings (SeEm), the PLM was then refined on general paraphrase identification (PI) datasets to learn the matching task (Step 2, Figure 3). [12]

In this paper, the extended pretraining (Step 1, Figure 3) is considered on additional domain-specific literature. In addition, the evaluation dataset is extended to include properties from air handling units (Step 3, Figure 3). However, the focus of the extension of the first model design from [12] is the integration of meta-information of the properties as complementary parameters for a previous classification of the properties (Steps I-II, Figure 3).

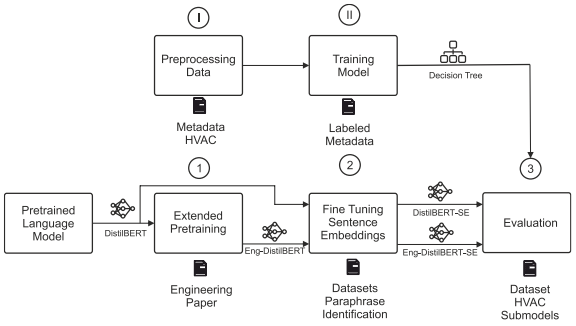

Fig. 3 – Training process

2. Background

Interoperability is a central pillar in the efforts to transform systems of automation technology into self-organizing systems [2]. It requires a digital representation of systems and the structured provision of their information. SI is required for mutual understanding of the information.

2.1 The Asset Administration Shell

For the digital representation of technical components, different concepts like [13] or [14] have been established in the technical domain. In this paper, we build on the concept of the asset administration shell [14] (AAS) as the digital representative of an asset. In this respect an asset can be understood as any entity (physical or logical) being of value to an enterprise [15]. The AAS originates from the field of Industrie 4.0 (I4.0), but can also be used for the digitalization of TBE. The

composition of asset and AAS is referred to as an I4.0-component [14].

A prerequisite for the automated exchange of information is the representation of all component information in standardized digital form within an AAS. This is realized by structuring the content into standardized submodels, which represent topics such as identification, design and configuration[14]. In the submodels, associated properties and functionalities of the asset are represented as submodel elements (SEs). SEs inherit mandatory and optional meta-level attributes from higher-level classes (Figure 4) [14].

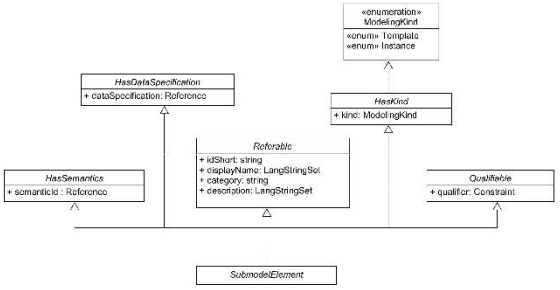

Fig. 4 - Class of the SE with its superordinate classes (own representation based on [14])

The class HasSemantics prescribes, for example, the mandatory definition of a SemanticID, which references a semantic definition of the SE. Further is the data type and the definition of the SE derivable as mandatory information from the class HasDataSpecification. [14]

2.2 Semantic Interoperability with homogenous semantics

To ensure interoperability according to the I4.0 research approach, the contents of the submodels must be standardized. Various initiatives standardize submodels for industrial components, e.g. pumps or drives[5, 8].

In addition to the uniform structure of the submodel contents by the information model, a uniform semantic language understanding of the information is required for the interoperable interaction of I4.0 components [4]. The integration of the submodels into dictionaries such as the ECLASS standard [16] enables the unambiguous identification of the submodels, as well as their SE, thus paving the way for homogeneous semantics.

Accordingly, the I4.0 approach requires a uniform structure (information model management shell) and semantics (homogeneous language space) to achieve interoperability. [14] Heterogeneous language spaces, i.e. language spaces that are not captured in standardized dictionaries, such as those available in manufacturer-specific descriptions, are not captured in the current I4.0 approach.

2.3 Semantic Interoperability with heterogeneous semantics

To achieve SI, it is possible to map heterogeneous semantics to each other in addition to using a uniform language space. This can be done either by semantically unique links to each other (Linked Data) or by automated matching.

Linked Data describes the use of ontologies to semantically describe entities. Ontologies represent knowledge structurally in machine-readable format. By using standardized ontological markup languages, entities of different ontologies can be linked together, creating a semantic network, e.g. [17]. In the context of I4.0, linking allows the recognition of SEs' semantics of one AAS by other AASs. The use of Linked Data in the technical domain requires knowledge of different ontologies and the creation of links between entities. In a comprehensive I4.0 value network, a large number of domain-specific ontologies of different components can be expected. Linking these requires a high level of analysis and engineering effort [18]. This complicates the use of Linked Data models.

Automated matching is another mapping possibility. This is pursued in this paper. The method used here is based on concepts of NLP and enables configuration-free translation of vocabularies. It requires neither a common semantic standard nor a manual linking of heterogeneous language spaces.

NLP is a subfield of artificial intelligence that enables computers to understand natural language (textual and phonetic) in order to perform tasks/actions based on it [11]. State of the art NLP models are based on the transformer architecture [19]. It is characterized by an encoder-decoder structure. The encoder is used for speech analysis and classification and the decoder for speech generation tasks. To cover more specific requirements in the areas of classification or generation, current models are usually characterized by either an encoder (e.g. [20]) or decoder structure (e.g. [21]). The model architecture for the semantic matching developed within this paper is based on an encoder architecture.

The training of current LMs is divided into pretraining and finetuning [22]. First, LM are pre-trained on large amounts of text and thus already learn a general understanding of the language. Based on this understanding, LM are adapted to specific tasks in a second training (finetuning) [23]. LM's language understanding is based on the representation of words in vectorized form, the word embeddings (WE). These WE are adapted during pretraining with different texts to represent the semantic meaning and relationships of the words. Thus, they are already capable of recognizing similarities or relationships, e.g., whether a word pair like Berlin and Germany has the same relationship as Madrid and Spain. In finetuning, the

WE are refined with specific data sets of a concrete use case. For automated matching of heterogeneous descriptions, this paper builds on the PLM DistilBERT [24].

2.4 The ISO-DistilBERT-SE Model

The general DistilBERT model was pretrained on English literature in the form of the BooksCorpus [25] and English Wikipedia [24]. Research shows that PLMs for domain-specific use cases achieve better results with a second phase of pretraining on specific literature [26]. Therefore, DistilBERT was extended with training on ISO standards [27–33] to form the ISO-DistilBERT model (Step 1, Figure 3) [12].

Subsequently, the model was used to be refined into ISO-DistilBERT-SE using data sets from the PI domain (Step 2, Figure 3). PI refers to the ability of a model to detect whether two sentences have the same meaning [34]. Here, the methodology of the Sentence BERT [35] model is adapted. The WE are combined into a Sentence Embedding (SeEm) with fixed dimension. Using cosine similarity, it is possible to check whether the SeEm of two sentences are similar and thus determine whether they are paraphrases [35]. The datasets used are the general datasets MultiNLI [36], STS-Benchmark [37] and QQP [38].

The model was then evaluated on a dataset created for this purpose (Step 3, Figure 3). For the ISO-DistilBERT-SE model, this dataset contains different definitions and names for SEs from the pump identification and design domain.

The ISO-DistilBERT-SE model achieves an accuracy of 94.33%. This shows that the use of the SeEm is an effective method to determine the similarity of SE based on the name and definition. Furthermore, an increase in accuracy of 2.46 percentage points was achieved by the extended pretraining [12].

3. Extension of the Semantic Matching Model

In order to be able to use the model in concrete application scenarios, such as extensive plant asset management, the evaluation dataset will be extended to include additional SEs and associated paraphrases from the field of air handling system (AHs) technology. In addition, the following model developments will be made: More domain-specific literature will be added to the extended pretraining. Metadata will be integrated as a classification feature to increase the accuracy of the model.

3.1 Extension of the evaluation data

The data set for the evaluation of the model is extended from SE of pumps to SE of an entire air handling system. Each component of the AHS, as well as the AHS as a whole, have up to twelve submodels, which are divided into the topics identification, design, maintenance, control, etc. [8] In total, the target vocabulary is thus extended from a pump's 39 SE to 427 SE of an AHS. For each SE, up to eleven paraphrases consisting of name and definition are created. Thus, a total of 1052 paraphrases for the 427 SEs of the target vocabulary are available for the evaluation of the model. The model is tested to match any name and definition of a SE from the paraphrase dataset to the matching SE of the 427 possible SEs in the target vocabulary. The larger the target vocabulary, the more difficult it is for the model to predict the matching SE of the target vocabulary from the possible SEs.

3.2 Extended pretraining on domain-specific literature

The extended pretraining from DistilBERT to ISO-DistilBERT is based on ISO standards [12]. This is complemented with another dataset on Eng-DistilBERT. The dataset includes 81.1M academic English papers from several disciplines [39]. These were filtered by engineering discipline, leaving 228,000 papers for the extended pretraining. Thus, the model learns complementary, subject-specific vocabulary of the engineering domain. The pre-trained model Eng-DistilBERT is then refined to the PI task following the presented procedure in [12] with the creation of SeEm. The refined model is referred to as Eng-DistilBERT-SE.

3.3 Extension of the model through metadata classification

The fault evaluation of the automated matching from [12] shows that errors occur mainly because different SE are often designated and defined very similarly. For example, there are several similarly defined pressure limits for a pump. For a more precise delimitation of the SEs among each other, a classification performed in advance is introduced in this paper. This is based on additional attributes, the metadata of the SE. Pressure limit values, which differ e.g., only by the time of their definition (definition during manufacturing vs. planning), can be distinguished from each other by an appropriate meta information and thus do no longer qualify as mutual paraphrase.

The first step is to select the metadata categories according to which a classifying algorithm delimits the SEs from each other. Since the whole model is based on the AAS information model, the metadata defined there are used for SE. The first metadata category *AAS_Spec* distinguishes the SE into the different subclasses property, file or SE collection. The properties, representing the majority of the SE, are further differentiated by the meta information of the *category* from the class Referable (Figure 4) according to the attribute Variable, Parameter or Constant. From the HasDataSpecification class (Figure 4), the *data type* and *unit* of an SE are taken as metadata categories. The *qualifier* according to

[41] is included as the last metadata to differentiate SEs according to their lifecycle status. The metadata selection is tailored to the present AHU plant evaluation dataset in order to characterize the existing SEs according to the most meaningful metadata possible. For this purpose, the metadata categories with their expressions in every possible combination are elaborated and labeled for a classification algorithm (Table 1). Since not all meta information is mandatory according to [14] and [41] e.g. the case that no qualifier is defined for an SE is also considered as a possible combination.

Tab. 1 – Label of possible metadata combinations

AAS_ Spec	Qualifier	Category	Unit_ Categ	Meta-label
File	OP	-	-	1
File	-	-	-	1|2
File	SUP	-	-	2

Thus, an SE can be labeled based on its metadata. The task of automated labeling places a multi-class requirement on a classification algorithm. For this purpose, a decision tree (DT) is trained according to [42] , which forms a structure of as few goal-directed decision paths as possible from the data set of possible metadata specifications in order to classify the data effectively. The points on a path where the next data decomposition is decided are called nodes. The tree structure of the "Decision Tree Classifier" according to [42] prescribes the maximum formation of two classes per node. Thus, questions are asked in a node, which can be answered binary with yes or no. The structure of the DT and the formation of the next node is determined automatically according to the best Gini-Impurity [42] calculation of the algorithm. The data basis of the DT are the possible metadata combinations and associated labels to be predicted (excerpt in Table 1). Figure 5 shows a section of the decision tree created.

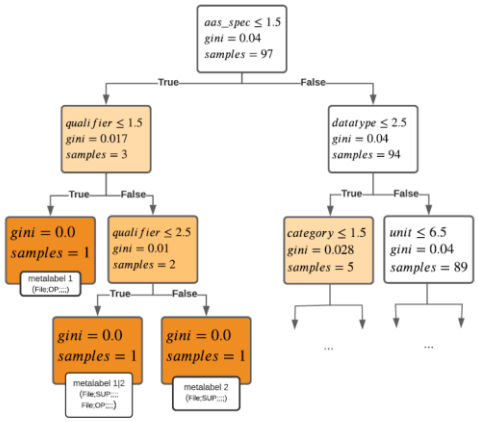

Fig. 5 - Detail of the decision tree

The metric used to evaluate the model is accuracy.

$$accuracy = \frac{correct\ predictions}{all\ predictions}$$

Training the DT on possible metadata combinations achieves 100% accuracy.

3.4 Combination of the Classifier and the LM Eng-DistilBERT-SE

The processing of the metadata of an SE is used for pre-filtering, indicating which SE of the target vocabulary should be reasonably used for comparison, in order to find the correct paraphrase. Thus, the SEs of the target vocabulary that cannot represent a paraphrase due to different metadata are dropped. By reducing the number of possible paraphrases, the probability of a wrong assignment is minimized.

Both the LM Eng-DistilBERT-SE and the DT are trained separately on their task to be solved. Eng-DistilBERT-SE processes the semantic information of an SE. The DT assigns one or more possible metadata classes to the SE. In Semantic Matching (SM), first, the SeEms and the metadata class(es) of the SE in the target vocabulary are formed and the information is linked. For unique identification, the SEs of the target vocabulary are indexed. In the use case, an unknown SE is labeled with metadata class(es) and then receives the indices of the SEs of the target vocabulary that qualify as paraphrases based on their metadata class. The LM Eng-DistilBERT-SE then compares the SeEm of the SE only with the SeEm of the SE that is in the indexed selection(Figure 6).

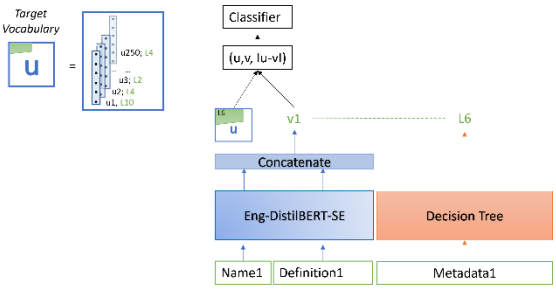

Fig. 6 - Advanced model MetaEng-DistilBERT-SE

Thus, the additional information of the metadata class(es) acts as a filter for the PI of the LM. The fused model is called MetaEng-DistilBERT-SE.

4. Evaluation

Table 2 presents the results of the different models on the newly created evaluation dataset. Increasing the target vocabulary from 39 to 427 decreases the accuracy of the original DistilBERT-SE model from 94% to 64.7%. The improvement of the accuracy by 2.3 percentage points of the model ISO-DistilBERT-SE to DistilBERT-SE in [12] already showed that pretraining with technical literature is promising. This is confirmed by the renewed increase in accuracy by another 2.5 percentage points by adding

engineering papers to the pretraining (Eng-DistilBERT-SE).

Using the integrated metadata model, the MetaEng-DistilBERT-SE model achieves an increase in accuracy of another 16.3 percentage points in the PI of SE task. This confirms the positive effect of classification by metadata in parallel with the processing of semantic information from SE to PI.

Tab. 2 - Semantic Matching results

Model version	Accuracy
DistilBERT-SE	64.7%
ISO-DistilBERT-SE	67.0%
Eng-DistilBERT-SE	69.5%
MetaEng-DistilBERT-SE	85.8%

Table 3 shows the influence of each metadata category on the model's output.

Tab. 3 - Analysis of different variants of the metadata model MetaEng-DistilBERT-SE

MetaEng-DistilBERT-SE	Acc
w/all Metadata	85.8%
w/o AAS_Spec	85.7%
w/o Datatype	85.6%
w/o Qualifier	85.4%
w/o Unit	79.9%
w/o Category	68.8%

The greatest influence on the model is the classification of the data according to the meta information of *category*. It differentiates SEs from each other by distinguishing variable from parameterized and constant values. This not only helps to distinguish measured values from nominal values, but also covers life cycle status. Values defined by the manufacturer in the design phase are defined as constant. In contrast, the designer's specifications are classified as parameters, and operational values are classified as variable. One of the main sources of error in the Eng-DistilBERT-SE model was the distinction between similarly defined features that differ only in their life cycle status. Therefore, solving this problem by defining the *category* is particularly influential on overall performance. Since the *qualifier* is only optionally present in the data, it has less significance as a category than the compulsorily defined *category* of the SEs. The *unit* category helps to differentiate measurement values among themselves whose metadata are not distinguished from each other by the other categories. The omission of one of the metadata categories *data type* and *AAS_Spec* exerts little influence on the model since the remaining categories of the SE already unambiguously determine the metadata class in the respective version. A constant SE of the dataset is e.g., in 98.5 % of the cases a SE of the datatype String or Real. Without specification of the data type these two classes are nevertheless distinguished by the specification of the *unit* with physical SE. Overall, the influence of individual categories for the classification of SE depends on the available data distribution. However, the clear improvement of the PLM can be statued by the inclusion of a metadata classification.

5. Conclusions

The results extend the current I4.0 research approach to the interaction of semantically homogeneous to semantically heterogeneous AAS. For this purpose, a combination of LMs, for mapping heterogeneous semantics, and a DT, for prior classification of the SE, was used. The results show that classification by the DT produces a significant increase in accuracy (70 to 86%). However, this needs to be further increased for acceptance in practice. For this purpose, other methods from the NLP field are applied to investigate how the accuracy can be increased. In addition, the information model of the AAS will be analyzed with respect to further metadata that can be used for classification.

Besides improving the model, it will be prototypically implemented as an I4.0 service in an I4.0 environment [43]. For this purpose, an I4.0 interface is specified that can be used within AAS to implement an SM service. The results are fed into relevant specification work to extend the interaction manager of AAS with an SM service.

The datasets generated during the current study are available in the Labor GART repository, https://github.com/thcologne-gart

6. Acknowledgement

The authors gratefully acknowledge financial support from the KSB Foundation in the project Automatic interaction of semantically heterogeneous Industrie 4.0 Asset Administration Shells by means of generic translation mechanisms based on methods of Natural Language Processing (1.1359.2020.1).

References

[1] Federal Ministry for Economic Affairs and Energy (BMWi). 2030 Vision for Industrie 4.0: Shaping Digital Ecosystems Globally; October 2019.

[2] Federal Ministry for Economic Affairs and Energy (BMWi). Position Paper: Interoperability – Our vision for Industrie 4.0: Interoperable communication between machines within networked digital ecosystems;

November 2019.
[3] International Electrotechnical Commission. Semantic Interoperability: Challenges in the digital Transformation Age: IEC Whitepaper 2019.
[4] VDI. Language for I4.0 components: Structure of messages. VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik (GMA); 2020 04.2020.
[5] ZVEI - Zentralverband Elektrotechnik und Elektronikindustrie e. V. Antrieb 4.0 – Vision wird Realität: Merkmale, Daten und Funktionen elektrischer Antriebssysteme in Industrie 4.0 für Hersteller, Maschinenbauer und Betreiber; November 2018.
[6] DIN, DKE, STANDARDIZATION COUNCIL INDUSTRIE 4.0. Deutsche Normungsroadmap Industrie 4.0: Version 4; März 2020.
[7] Federal Ministry for Economic Affairs and Energy (BMWi). Submodel Templates of the Asset Administration Shell: ZVEI Digital Nameplate for industrial equipment (Version 1.0); November 2020.
[8] Both M, Müller J. Digitization of pumps – Industry 4.0 submodels for liquid and vacuum pumps. 4th International Rotating Equipment Conference 2019.
[9] Ostermeier M, Maisch N, Both M, Ulmer R, Müller J. BIM im Betrieb durch lebenszyklusübergreifende Verfügbarkeit von Anlagendaten auf Basis von I4.0-Verwaltungsschalen. In: Automation 2021. VDI Verlag 2021; 131–40.
[10] Clemens Rohde, Patrick Plötz, Lisa Nabitz, *et al.* Branchen- und unternehmensgrößenbezogene Ermittlung von Klimaschutzpotenzialen (Schwerpunkt KMU) durch verstärkte Umsetzung von Energiemanagementmaßnahmen in der Wirtschaft: Abschlussbericht; 2018.
[11] Jurafsky D, Martin JH. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. 2. ed., Pearson internat. ed. Upper Saddle River, NJ: Prentice Hall Pearson Education Internat 2009.
[12] Both M, Müller J, Diedrich C. Automatisierte Abbildung semantisch heterogener I4.0-Verwaltungsschalen durch Methoden des Natural Language Processing. at - Automatisierungstechnik 2021; 69(11): 940–51 [https://doi.org/10.1515/auto-2021-0050]
[13] Kaebisch et al. Web of Things (WoT) Thing Description; 2020 [cited 2021 November 13] Available from: URL: https://www.w3.org/TR/wot-thing-description/.
[14] Federal Ministry for Economic Affairs and Energy (BMWi). Details of the Asset Administration Shell: Part 1 - The exchange of information between partners in the value chain of Industrie 4.0 (Version 2.0.1). Berlin; May 2020.
[15] VDI/VDE-Gesellschaft. Industrie 4.0 Begriffe/Terms: VDI Statusreport; April 2019.
[16] ECLASS. ECLASS - Standard für Stammdaten und Semantik für die Digitalisierung [cited 2020 November 17] Available from: URL: https://www.eclass.eu/index.html.
[17] Linked Open Data Cloud. The Linked Open Data Cloud [cited 2020 November 16] Available from: URL: https://lod-cloud.net/.
[18] Automated Ontology Matching in the Architecture, Engineering and Construction Domain - A Case Study. Schneider, Georg. The 7th Linked Data in Architecture and Construction Workshop 2019.
[19] Vaswani A, Shazeer N, Parmar N, *et al.* Attention Is All You Need; 2017 Jun 12.
[20] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding; 2018 Oct 11.
[21] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners Alec; 2019.
[22] Ruder S, Peters ME, Swayamdipta S, Wolf T. Transfer Learning in Natural Language Processing. In: Sarkar A, Strube M, editors. Transfer Learning in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics; 15–8.
[23] Peters ME, Ruder S, Smith NA. To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks; 2019 Mar 14.
[24] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter; 2019 Oct 2.
[25] Zhu Y, Kiros R, Zemel R, *et al.* Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books; 2015 Jun 22.
[26] Gururangan S, Marasović A, Swayamdipta S, *et al.* Don't Stop Pretraining: Adapt Language Models to Domains and Tasks; 2020 Apr 23.
[27] ISO. ISO TC 115: Pumps [cited 2020 December 11] Available from: URL: https://www.iso.org/committee/51766.html.
[28] ISO. ISO/IEC JTC 1/SC 17: Cards and security devices for personal identification [cited 2020 December 11] Available from: URL: https://www.iso.org/committee/45144.html.
[29] ISO. ISO/IEC JTC 1/SC 31: Automatic identification and data capture techniques [cited 2020 December 11] Available from: URL: https://www.iso.org/committee/45332.html.
[30] ISO. ISO/TC 184/SC 4: Industrial data [cited 2020 December 11] Available from: URL: https://www.iso.org/committee/54158.html.
[31] ISO. ISO/TC 46/SC 4: Technical interoperability [cited 2020 December 11] Available from: URL: https://www.iso.org/committee/48798.html.
[32] ISO. ISO/TC 46/SC 9: Identification and description [cited 11.12.20] Available from:

URL:
https://www.iso.org/committee/48836.html.
[33] ISO. ISO/TC 59/SC 2: Terminology and harmonization of languages [cited 11.12.20] Available from: URL:
https://www.iso.org/committee/49076.html.
[34] Socher R, Huang EH, Pennington J, Ng AY, Manning CD. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In: Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection; 2011. Red Hook, NY, USA: Curran Associates Inc; 801–9.
[35] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks; 2019 Aug 27.
[36] Williams A, Nangia N, Bowman S. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In: A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference; 2018. Association for Computational Linguistics; 1112–22.
[37] Cer D, Diab M, Agirre E, Lopez-Gazpio I, Specia L. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In: SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation; 2017. Vancouver, Canada: Association for Computational Linguistics; 1–14.
[38] Iyer S, Dandekar N, Csernai K. First Quora Dataset Release: Question Pairs: Quora [cited 2020 November 23] Available from: URL:
https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs.
[39] Lo K, Wang LL, Neumann M, Kinney R, Weld D. S2ORC: The Semantic Scholar Open Research Corpus. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. S2ORC: The Semantic Scholar Open Research Corpus. Stroudsburg, PA, USA: Association for Computational Linguistics; 4969–83.
[40] DIN Deutsches Institut für Normung e.V. DIN EN 62569-1 Allgemeine Regeln zur Erstellung von Produktspezifikationen: Teil 1: Grundsätze und Methoden; 2018.
[41] Pedregosa et al. Scikit-learn: Machine Learning Library for Python; 2011.