MaintIE: A Fine-Grained Annotation Schema and Benchmark for Information Extraction from Maintenance Short Texts

Tyler Bikaun, Tim French, Michael Stewart, Wei Liu and Melinda Hodkiewicz
The University of Western Australia
Perth, Western Australia, Australia
{tyler.bikaun@research., tim.french@, michael.stewart@, wei.liu@, melinda.hodkiewicz@}uwa.edu.au

Abstract

Maintenance short texts (MST), derived from maintenance work order records, encapsulate crucial information in a concise yet information-rich format. These user-generated technical texts provide critical insights into the state and maintenance activities of machines, infrastructure, and other engineered assets–pillars of the modern economy. Despite their importance for asset management decision-making, extracting and leveraging this information at scale remains a significant challenge. This paper presents MaintIE, a multi-level fine-grained annotation scheme for entity recognition and relation extraction, consisting of 5 top-level classes: PhysicalObject, State, Process, Activity and Property and 224 leaf entities, along with 6 relations tailored to MSTs. Using MaintIE, we have curated a multi-annotator, high-quality, fine-grained corpus of 1,076 annotated texts. Additionally, we present a coarse-grained corpus of 7,000 texts and consider its performance for bootstrapping and enhancing fine-grained information extraction. Using these corpora, we provide model performance measures for benchmarking automated entity recognition and relation extraction. The MaintIE scheme, corpus, and model are publicly available at `https://github.com/nlp-tlp/maintie` under the MIT license, encouraging further community exploration and innovation in extracting valuable insights from MSTs.

Keywords: Information Extraction (IE), Named Entity Recognition (NER), Relation Extraction (RE), Maintenance Work Orders (MWO)

1. Introduction

Natural language processing (NLP) has advanced significantly, with numerous datasets, such as ACE (Doddington et al., 2004), CoNLL03 (Tjong Kim Sang and De Meulder, 2003) and others, enabling automated text analysis. However, in the specific area of 'engineering maintenance', which deals with the state and upkeep of engineered objects, there is a notable scarcity of specialised datasets and models (Brundage et al., 2021; Dima et al., 2021).

Maintenance is an indispensable cornerstone in today's industrialised world, profoundly impacting diverse sectors, including defence, energy, manufacturing, infrastructure, resources, and transportation. It accounts for billions of dollars in operational expenditure globally annually (Thomas, 2018; Brundage et al., 2019). Within this essential domain, the short text field of 'maintenance work order' (MWO) records serves as a key medium for capturing information about engineered objects, which we refer to as maintenance short text (MST). These are concise, user-generated texts, similar in brevity and quality to a Tweet but with a medical note's complexity and technical nature (Figure 1) (Brundage et al., 2021). The critical nature of these texts lies in their support for domain experts in various analytical tasks, such as equipment failure analysis (Lee et al., 2023), root cause investigations (Valcamonico et al., 2024), and the development of performance indicators (Lukens et al., 2019).

Like semantic role labelling (Gildea and Jurafsky, 2002), the significance of information extraction from MSTs hinges on deciphering "*who* is doing *what* to *what*, and *why*?". Within this context, the 'who' may refer to either an engineered object or a person, the 'what' pertains explicitly to an engineering object and the 'why' reveals the condition or state that necessitates the action on the entity. Take the example: "replace the blown air conditioner motor". In this case, the extracted information would ideally discern the following facts: (i) the *air conditioner* has a part called the *motor*, (ii) the *motor* is subject to the *replace* activity, and (iii) the *motor* has the *state* of being *blown*. This type of semantic comprehension is crucial for the practical analysis and use of MSTs, as it explicitly identifies the object, its parts, their condition and any required actions identified in the MWO record.

Despite the significance of maintenance to the economy, exploring MSTs from an information extraction perspective remains under-researched largely due to the limited publicly available annotated datasets (Dima et al., 2021; Stewart et al., 2022). Addressing this gap, this paper presents **MaintIE** (**Maint**enance **I**nformation **E**xtraction) – a fine-grained scheme and two annotated corpora for information extraction of MSTs. MaintIE focuses on entity recognition and relation extraction of low-quality MSTs that have been lexically normalised. MaintIE is publicly available to facilitate

LREC-COLING 2024, pages 10939–10951

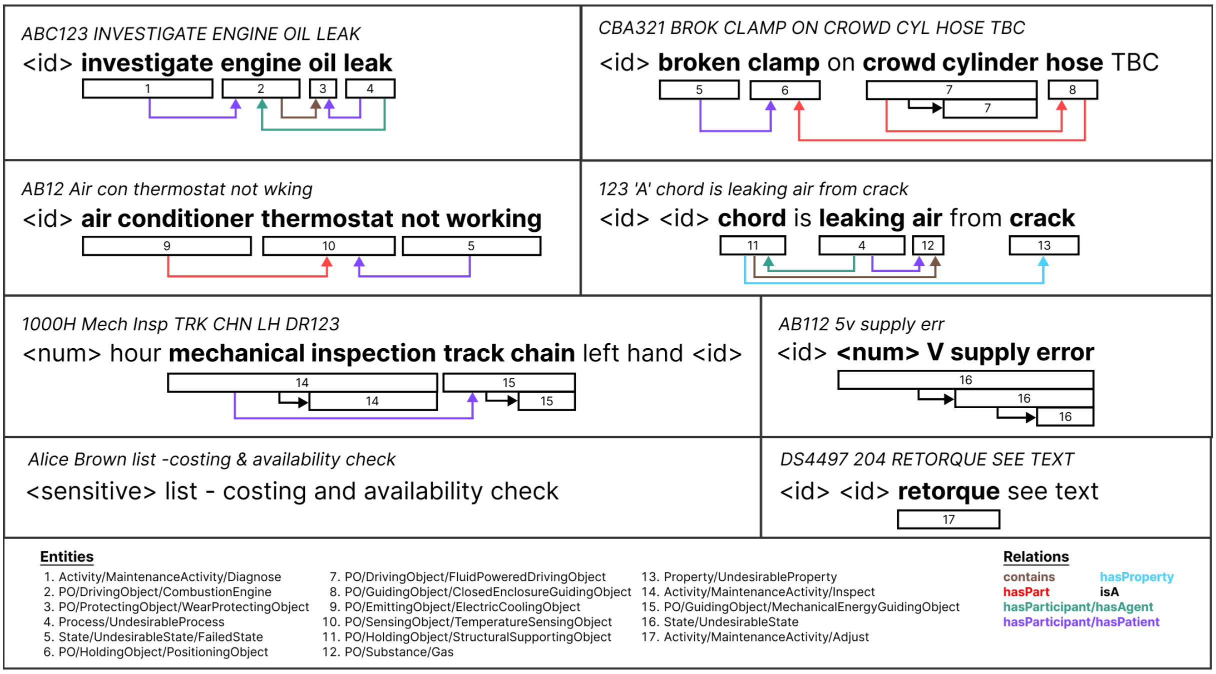

Figure 1: An illustration of the MaintIE annotation scheme applied to lexically normalised maintenance short texts. Italicised portions represent the original text, which is transformed into the larger text through lexical normalisation and sanitisation. The entity prefix 'PO' indicates a *PhysicalObject*.

reproducible research and IE on MSTs[1].

The remainder of this paper describes related work (section 2), annotation scheme (section 3), dataset selection and preprocessing (section 4), annotation process (section 5), benchmark models trained and evaluated using the annotated MaintIE corpora (section 7), and finally, a discussion of the results (section 8).

2. Related Work

Engineering-specific texts, particularly in our area of focus – maintenance short texts (MST), have been used in NLP tasks such as text classification, sentiment analysis, knowledge representation and clustering (Stenström et al., 2015; Saetia et al., 2019; Yang et al., 2020; Ottermo et al., 2021; Bhardwaj et al., 2021; Usuga Cadavid et al., 2022; Iyer et al., 2022). Existing research addressing IE for MSTs has been limited to Named Entity Recognition (NER). For example, the works of Sexton et al. (2018); Navinchandran et al. (2021) introduced a three-class NER system, distinguishing entities as *Problem*, *Item*, and *Solution*. Similarly, Gao et al. (2020) implemented a three-class NER system based on *Item*, *Action*, and *State*, echoing the classification approach of the prior authors. Advancing from these foundational studies, Stewart et al. (2022) expanded the categorisation with a ten-class NER system, broadening the scope to *Activity*, *Agent*, *Attribute*, *Cardinality*, *Consumable*, *Item*, *Location*, *Observation*, *Specifier*, and *Time*. Using a pretrained bidirectional short-term memory (LSTM) model with character-level embeddings, Stewart et al. (2022) achieved an impressive 82.8% micro F1 on a dataset comprising 4,000 annotated entities on MSTs.

While the named entities in existing works can identify the "who," "what," and "why" in MSTs, they do not provide a complete semantic comprehension of the complex dynamics within these texts. To answer questions like "who is doing what to what, and why?" we need to extract inter-entity relations (Jurafsky, 2000), which is the focus of relation extraction (Pawar et al., 2017). However, relation extraction for MSTs is still in its early stages, with the only notable attempt being one by Stewart et al. (2022), which employs unsupervised co-occurrence-based heuristics for generating inter-entity relationships. Nevertheless, there is a growing interest in constructing semantic graphs from technical engineering-specific texts (Sarica et al., 2020; Bhardwaj et al., 2023; Wang and El-Gohary, 2023); however, these are yet to be applied to MSTs.

There are several significant challenges when it comes to IE on MSTs. Firstly, there is a lack of publicly available annotated IE datasets, with only one corpus being released by Stewart et al. (2022). This is because MWOs are often confidential, and organisations are hesitant to release their data (Akhbardeh et al., 2020; Bikaun et al.,

[1] https://github.com/nlp-tlp/maintie

2024). Additionally, human annotation of technical texts like MSTs is expensive and requires significant subject-matter expertise and tacit knowledge (Settles, 2009). Secondly, MSTs often have poor quality language (Stenström et al., 2015; Hodkiewicz and Ho, 2016; Akhbardeh et al., 2020; Bikaun et al., 2024), requiring subject-matter expertise for text normalisation (Bikaun et al., 2021). Addressing pre-processing and quality issues is essential to ensure optimal performance of downstream tasks (van der Goot et al., 2021). Lastly, the existing entity schemes used for MSTs can be suboptimal for IE since they are shallow and rely on general classifications such as "Item" and "Activity" to categorise nouns and verbs (Sexton et al., 2018; Gao et al., 2020; Navinchandran et al., 2021; Stewart et al., 2022). This overuse of general classification results in a loss of specificity, leading to challenges when disambiguating texts in downstream engineering applications.

3. Annotation Scheme

This section outlines the MaintIE annotation scheme, with its entity classes and relations briefly described in Table 1. In developing the scheme, we are guided by eight design requirements outlined in the repository. These requirements were established to ensure that the scheme is efficient, expressive, and easily understood by all stakeholders – ranging from annotators to the domain experts who engage with the scheme and its outputs. Further details pertaining to the MaintIE scheme can be found by consulting the repository.

Entities MaintIE features 224 distinct entity classes across 3 levels, organised under 5 primary hierarchical categories: *PhysicalObject*, *State*, *Process*, *Activity*, and *Property*. An illustrative depiction of these classes is available in Figure 1. MaintIE adopts a systems engineering approach to annotation of physical objects utilising the *inherent function* of objects as the basis for item classification (Stone and Wood, 1999). This allows for directly using the function-based object classification section in the IEC 81346-2:2019 standard (IEC, 2019). Moreover, MaintIE borrows from established ontologies for state and activity classifications (Woods et al., 2021, 2023).

Relations MaintIE presents four distinct categories of relations: *mereological*, *property*, *type*, and *participatory*. Participatory relations are inspired by PropBank's (Kingsbury and Palmer, 2002) agent (ARG0) and patient (ARG1) labels accessed via the Unified Verb Index[2]. These labels are semantically defined as Proto-Agent properties (Dowty, 1991). As MaintIE uses a closed set of typed entities, these labels have been transformed into the semantic relations *hasParticipant/ hasAgent* and *hasParticipant/ hasPatient* to capture the participation of entities as agents, causes, or experiencers, as well as those undergoing a change of state or being affected by actions. For a visual representation of these relations, please refer to Figure 1.

4. Dataset and Preprocessing

The dataset used for MaintIE contains 10,000 MSTs randomly sampled from MWOs on assets used in the heavy mobile equipment, rail, mineral processing and infrastructure industries (Figure 2A). This random sampling exposes the scheme to a range of heavy-industry maintenance contexts. The domain-expert annotators, who are the first and last authors, also have direct work experience in these industries.

As described in the section 2, MSTs texts have poor lexical quality and also make considerable use of abbreviations and acronyms that are often unique to the company or maintenance context. The open-source annotation tool LexiClean (Bikaun et al., 2021) was used by a domain expert familiar with lexical normalisation and industrial maintenance (first author) to lexically normalise and sanitise the MaintIE corpus following the guidelines outlined by Bikaun et al. (2024). Confidential and non-semantic information is masked with the special tokens <id>, <sensitive>, <num> and <date>, while non-standard words and phrases are normalised into their canonical forms. Following this process, the texts are tokenized based on white space before semantic annotation. Figure 1 illustrates the outputs of this transformation process.

5. Annotation Process

This section outlines the process of annotating entities and relations in the MaintIE dataset (see Figure 2A). The task was performed by two domain experts, the first and last authors, who have over 30 years of combined experience in engineering and maintenance and are knowledgeable about entity and relation annotation. They used QuickGraph (Bikaun et al., 2022), a freely available tool, for the annotation task.[3] The annotators identify textual

[2]Unified Verb Index. https://verbs.colorado.edu/verb-index/index.php

[3]QuickGraph. https://quickgraph.nlp-tlp.org

Entity Type (Size)	Brief Description	Example
Activity $(1 \rightarrow 2 \Rightarrow 20)$	Activities related to maintenance and support tasks performed on physical objects, based on Woods et al. (2023).	*change out*, *overhaul*, *jump-start*
PhysicalObject $(1 \rightarrow 19 \Rightarrow 173)$	Function-based classification of physical objects based on IEC 81346-2:2029 (IEC, 2019).	*pump*, *engine*, *gasket*
Process $(1 \rightarrow 2)$	Temporal events relating to physical object function or condition.	*leaking*, *weeping*
Property $(1 \rightarrow 2)$	Attributes of a physical object, either essential (intrinsic) or accidental (non-intrinsic) (Robertson and Atkins, 2013).	*crack*, *hole*, *pressure*, *temperature*
State $(1 \rightarrow 2 \Rightarrow 3)$	Conditions of physical objects based on (Woods et al., 2021).	*blown*, *seized*
Relation Type	**Brief Description**	**Example**
contains	Mereological: Denotes containment of physical objects.	engine *contains* oil
hasPart	Mereological: Represents part-whole relationships of physical objects.	engine *has part* radiator
hasParticipant/hasAgent	Participatory: Represents entities actively involved or initiating an action or event. Equivalent to PropBank agent (ARG0) role (Kingsbury and Palmer, 2002).	leak *has agent* pump
hasParticipant/hasPatient	Participatory: Denotes entities passively affected by or undergoing an action or event. Equivalent to PropBank patient (ARG1) role (Kingsbury and Palmer, 2002).	leak *has patient* water
hasProperty	Property: Indicates the possession of a particular characteristic or attribute by an entity.	pipe *has property* temperature
isA	Type: Describes a subtype or instance relationship between entities.	diesel engine *is a* engine

Table 1: An overview of the entities and relationships within the MaintIE scheme. For in-depth details, consult the repository. *Size* represents the hierarchical structure for each top-level entity class.

spans corresponding to entities and assign a hierarchical entity type to each. If applicable, relations between the annotated entities are then applied. An example item from the annotated MaintIE corpus is shown in Appendix 10.1.

The project was allotted 6 weeks for annotation by the two subject-matter experts, during which 3,000 texts were randomly sampled from the 10,000 texts to be annotated. Both annotators worked independently using the entire 224 entity classes and 6 relations, meeting for two hours weekly to resolve issues and develop annotation guidelines. The agreement between the annotators was measured through the F1 score (van Rijsbergen, 1979) on entities, strict relations, and loose relations. An entity agreement was counted when both annotators agreed on the entity's span and type. A strict relation agreement required both annotators to recognise the identical relation type and match both the head and tail entity types and spans, while a loose relation agreement was reached when both annotators aligned on the relation type and the spans of the head and tail entities. Only texts that had perfect agreement were included in the final expert-annotated corpus. After 6 weeks, 1,400 texts were double annotated, and 1,067 achieved 100% agreement. These texts are included in the "Fine-Grained Expert-Annotated" gold-standard MaintIE corpus. The entire process took approximately 60 total hours, averaging 2.5 minutes per text.

The Fine-Grained Expert-Annotated corpus was used to train a deep learning model (section 7). The output of this model "pre-annotated" the remaining 7,000 texts (see Figure 2A), resulting in a second, coarse-grained annotated corpus with only the 5 top-level entity classes and 6 relations. To complete this process, we used the SpERT model (Eberts and Ulges, 2020), which was trained on the 5 top-level classes of the fine-grained corpus. We opted for this approach due to the scarcity of resources for annotation and the time-consuming nature of annotating the full-depth hierarchy. The purpose of the coarse-grained corpus is to determine whether using it as an additional training resource produces better fine-grained results due to enhanced entity and relation identification.

A single domain expert (first author) then used QuickGraph to correct residual annotation errors, resulting in the "Coarse-Grained Large-Scale" corpus. The coarse-grained annotation process on the pre-annotated corpus was much faster, taking approximately 40 hours to complete, due to the significantly fewer entity classes involved, averaging 25 seconds per text. Future work will explore using crowd-sourcing or equivalent to increase the size of this corpus with finer-grained entity classes.

6. Annotated Corpora Statistics

This section presents a summary of the statistics for both annotated MaintIE corpora. A total of 8,076 texts were annotated, encompassing 43,674 tokens with a vocabulary size of 2,409. The distribution of tokens per text ranged from a minimum of 1 to a maximum of 13, with an average of 5.4. Table 2 summarises the statistics for both corpus portions, while Appendix 10.2 details the entity and relation distributions across each.

6.1. Fine-Grained Expert-Annotated

The Fine-Grained Expert-Annotated corpus comprises 1,067 texts with 100% annotator agreement on both entities and relations. MaintIE fine-grained has 3,397 annotated entities (301 unique) with an average number of 3.2 entities per text. This accounts for 55.6% of the corpus tokens being assigned an entity. The number of relations is lower than that of entities, likely because relations require two entities to exist. A total of 2,341 relations (1,757 unique) were applied, with 38.3% of tokens participating in a relation. All 1,076 texts have at least two entities and one relation.

6.2. Coarse-Grained Large-Scale

The Coarse-Grained Large-Scale corpus consists of 7,000 texts pre-annotated using a deep learning model trained on the fine-grained corpus (see Figure 2A). The annotations were manually validated by a single annotator (first author) without inter-annotator agreement metrics to report. The corpus was annotated at the coarsest entity level, using only the 5 top-level entity classes and the same relations used in the fine-grained corpus. This coarse-grained corpus contains 22,122 entities (3,298 unique) and 15,200 relations (12,898 unique).

7. Experiments and Evaluation

7.1. Task Description

The main objective of MaintIE is to extract entities ($E = \{e_1, ..., e_n\}$) and relations ($R = \{r_1(e_n, e_m), ..., r_n(e_n, e_m)\}$) expressed in an MST with both coarse and fine-grained annotations. In our experiments, we explore the performance of token classification (using SpERT (Eberts and Ulges, 2020)) and sequence-to-sequence methods (using REBEL (Cabot and Navigli, 2021)). The aim of both methods is to accurately identify and assign the appropriate hierarchical entity type to textual spans, as well as recognise and classify inter-entity relations. In this context, relations can exist between any entity, and entities can be nested at any arbitrary depth. To evaluate the performance of this task, we use micro and macro-F1 scores measured on entities and relations (loose and strict) following Eberts and Ulges (2020) and Cabot and Navigli (2021).

7.2. Models

SpERT *Span-Based Joint Entity and Relation Extraction with Transformer Pre-Training* (SpERT) (Eberts and Ulges, 2020) operates as a token-classification model designed to extract entities and relations jointly. For this study, our implementation integrates SpERT with BERT (*bert-base-cased*) (Kenton and Toutanova, 2019). We adhered to the default hyperparameters, with the sole exception being the maximum span length, adjusted to 32.

REBEL The *Relation Extraction By End-to-end Language generation* (REBEL) (Cabot and Navigli, 2021) model extracts entities and relations as triples from text using an autoregressive sequence-to-sequence mechanism. This approach encodes semantic triples into a reversible linearised structure. REBEL is pre-trained over BART (*facebook/bart-large*) (Lewis et al., 2020) for relation extraction from a large subset of Wikipedia articles. Despite its exceptional performance on benchmark datasets, REBEL has a drawback in that it cannot extract entities without relations; hence, we only evaluate this model on its RE performance. For the most part, we have kept the default hyperparameters but have made specific adjustments to the maximum source and target lengths, setting them to 64.

7.3. Experimental Setup

We evaluated each model's ability to extract entities and relations with different hierarchies, ranging from coarse to fine, and tested their performance using the two annotated (fine and coarse-grained) corpora. Our primary goal is to produce a performance baseline for information extraction on MSTs. A secondary goal is to see if using the coarse-grained corpus improves the model's fine-grained extraction performance. Establishing a baseline for information extraction performance across different levels of entity hierarchy enables us to compare MaintIE against existing works.

To conduct the experiments, we used two model training strategies (illustrated in Figure 2): fine-tuning on the fine-grained corpus (FG Fine-Tune) and pre-fine-tuning on the coarse-grained corpus followed by fine-tuning on the fine-grained corpus (CG Fine-Tune $\rightarrow$ FG Fine-Tune). We studied the effect of entity hierarchy on extraction, ranging

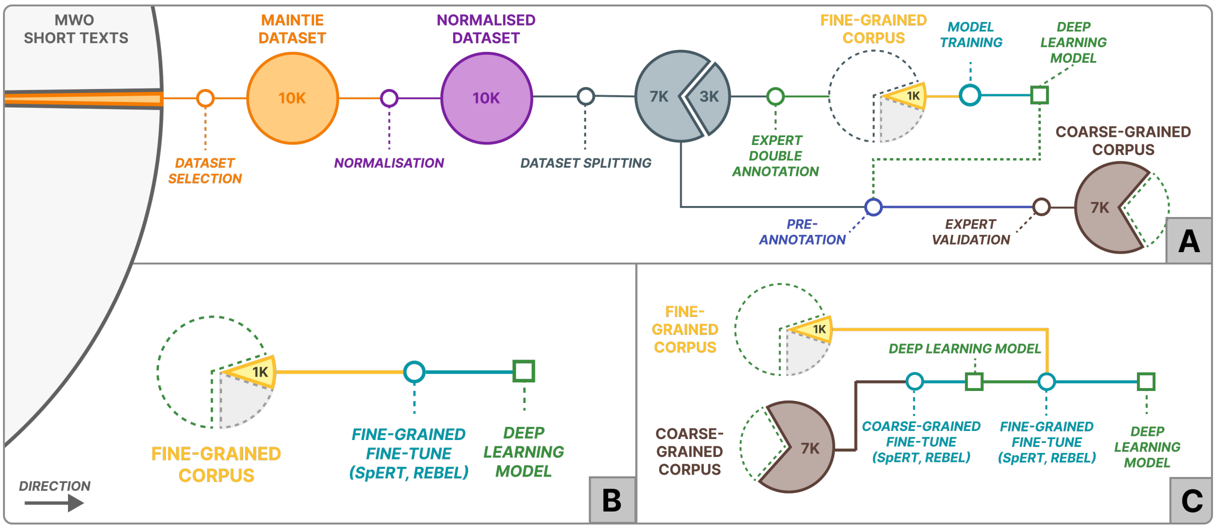

Figure 2: Workflows for MaintIE corpora creation (top) and experimental procedures (bottom): (A) Steps for developing Expert-Annotated and Large-Scale corpora. (B) Fine-tuning using only the Expert-Annotated corpus. (C) Fine-tuning using both MaintIE corpora.

from untyped[4], the coarse-grained level (5 classes), the intermediate level (32 classes), and the finest-grained level of 224 classes. For model training, development and testing, we used an 80/10/10 split on both corpora. Table 4 summarises the experimental results. Further details of the training strategies can be found in the repository.

8. Discussion

In this section, we explore the key outcomes of implementing our annotation scheme, the development of the two corpora, and the performance of the benchmark models.

8.1. Annotation Scheme and Corpora

The key distinguishing features of the MaintIE entity scheme are its multi-level structure, alignment with international standards for classifying engineered objects (IEC, 2019), and re-use of established ontologies for state and activities (Woods et al., 2023). In practice, annotating entities with a fine-grained corpus (of 224 classes) presents certain challenges, particularly with the bottom-level subclasses of *State* and *PhysicalObject*. This is due to the terse nature of the texts – 5.4 tokens on average. These texts often lack the context to disambiguate and situate named entities (e.g. 'cable' in 'replace cable' – is this a guiding or structural physical object?). Annotators often have to cross-check with external resources and reference material to corroborate correct entity classes. In the future, fortuitous information accompanying MSTs (Brundage et al., 2021) could be used to contextualise the texts and improve the annotation process. These issues are less pronounced when annotating the coarse-grained corpus, as only the top 5 levels of entity classes were used, which are easily differentiable and less ambiguous.

Of the 224 possible fine-grained entity classes, only 69.6%, or 156, were used in this corpus, primarily due to the comprehensive nature of the IEC 81346 functional hierarchy. This standard, intended to encompass all engineered assets, extends beyond the scope of our dataset, which is focused on mining operations' assets. This focus has resulted in a skewed distribution of physical object entities (illustrated in Appendix 10.2). This phenomenon presents challenges for task-specific learning by limiting the diversity of entities represented. Such a distribution underscores the need for corpus refinement to reflect better the varied and complex nature of maintenance activities across different industries.

Expanding the MaintIE corpus to include maintenance contexts from a broader range of engineered systems, such as those in the building or electronics industries, promises to increase its representativeness and utility substantially. This diversification strategy not only aims to address the noted limitations but also enriches the corpus, making it a more versatile resource for research and application in the wider field of maintenance engineering. Including diverse maintenance scenarios could provide a more holistic view of engineered asset management, significantly advancing the corpus towards achieving greater universality and relevance.

The MaintIE scheme differs from previous work in the domain due to its unique set of relations.

[4] one class (*Entity*) for SpERT and two classes (<*subj*>, <*obj*>) for REBEL.

Table 2: Summary statistics of the two annotated MaintIE corpora. *Density* refers to the number of tokens with the specified annotation type. *Schema utilisation* refers to the proportion of the schema applied to the dataset.

	Fine-Grained Expert-Annotated (1,076 texts)		Coarse-Grained Large-Scale (7,000 texts)	
	Entity	**Relation**	**Entity**	**Relation**
Total	3,397	2,341	22,122	15,200
Min. / Text	2	1	0	0
Max. / Text	6	6	7	7
Ave. / Text	3.2	2.2	3.2	2.2
Median / Text	3	2	3	2
Texts w/ Annotations	100%	100%	99.7%	96.1%
Annotation Density	55.6%	38.3%	58.9%	40.5%
Schema Utilisation	69.6% (156/224)	100% (6/6)	100% (5/5)	100% (6/6)

Table 3: Summary of entity and relation annotations in the annotated MaintIE corpora. Entity types are consolidated into their top-level classes for succinctness. For in-depth details, consult the repository.

		Fine-Grained Expert-Annotated				Coarse-Grained Large-Scale			
		Total		*Unique*		*Total*		*Unique*	
#	**Entity Type**	**Count**	%	**Count**	%	**Count**	%	**Count**	%
1	Activity	784	23.1	27	9.0	5,045	23.0	373	11.0
2	PhysicalObject	1,994	58.7	222	73.7	13,472	61.0	2,379	72.0
3	Process	146	4.3	9	3.0	728	3.0	118	4.0
4	Property	35	1.0	1	0.3	130	1.0	32	1.0
5	State	438	12.9	42	14.0	2,747	12.0	396	12.0
	Total	3,397	100	301	100	22,122	100	3,298	100
		Total		*Unique*		*Total*		*Unique*	
#	**Relation Type**	**Count**	%	**Count**	%	**Count**	%	**Count**	%
1	contains	38	1.6	15	0.9	178	1.2	137	1.1
2	hasPart	533	22.8	417	23.7	3,837	25.2	3,290	25.5
3	hasParticipant/hasAgent	166	7.1	127	7.2	789	5.2	716	5.6
4	hasParticipant/hasPatient	1,206	51.5	936	53.3	7,761	51.1	6,745	52.3
5	hasProperty	34	1.5	28	1.6	123	0.8	116	0.9
6	isA	364	15.5	234	13.3	2,512	16.5	1,894	14.7
	Total	2,341	100	1,757	100	15,200	100	12,898	100

Through our relation annotation, 17,541 relations were created across both MaintIE corpora. This effort has produced a comprehensive set of relations for MSTs and is the only human-annotated, freely available set in this domain. Relation annotation is challenging, especially for the fine-grained corpus, as getting an entity annotation wrong means the relation is also wrong when measured strictly. The original 1,400 double annotated texts had inter-annotator agreement scores, as measured by F1 score, of 91.8% for entities, 64.1% for strict relations, and 75.5% for loose relations. These less than 100% scores reflect the lack of agreement on 333 texts, as the remainder (1,067) have a perfect agreement. We decided to discard texts with non-perfect agreement from the corpus as they may have introduced inconsistency. Moreover, only annotating 1,400 of the original 3,000 texts allotted was due to the time limits on the annotation process.

During the annotation process, applying *mereological*, *property*, and *type* inter-entity relations was generally straightforward. However, most relation annotations, approximately 60%, as shown in Table 3, are participatory relations, which can present challenges. It is sometimes difficult to identify entities participating in the patient and agent roles. Many verbs and verb phrases are domain-specific and not captured in resources like VerbNet or PropBank. Additionally, it is challenging to disambiguate the main participants of the short texts even when the PropBank roleset is identified, as these texts lack proper grammatical structure and are concise. For example, in annotating the nouns “pulley" and “lagging" in “change out pulley lagging" and “change out pulley - lagging worn". Annotator agreement

Table 4: Results of automatic information extraction, represented as F1 scores. **Bold** text indicates the overall best-performing scores on the 224 entity classes. Underlines indicate the best score for the given model and training strategy. Abbreviations: *ER* (entity recognition), *RE* (relation extraction), *FG* (Fine-Grained Expert-Annotated corpus), and *CG* (Coarse-Grained Large-Scale corpus). For comprehensive results of the top-performing models, refer to Appendix 10.3.

Model	Fine Tuning Strategy	Entity Classes	ER		RE Strict		RE Loose	
			Micro	*Macro*	*Micro*	*Macro*	*Micro*	*Macro*
REBEL	FG (Figure 2B)	2	Not Relevant		68.9	69.7	68.9	69.7
		5			67.9	57.3	68.3	57.4
		32			0.8	0.6	69.0	70.2
		224			0.0	0.0	66.7	53.5
	CG → FG (Figure 2C)	2	Not Relevant		73.7	73.7	73.7	73.3
		5			69.0	70.9	70.7	71.1
		32			37.4	51.1	71.1	72.5
		224			0.0	0.0	**76.1**	**77.3**
SpERT	FG (Figure 2B)	1	89.9	89.9	72.9	74.7	72.9	74.7
		5	87.4	88.8	67.1	62.3	68.3	62.6
		32	71.6	58.9	36.5	42.8	65.5	61.9
		224	**64.4**	**38.7**	**27.3**	**35.1**	55.7	61.1
	CG → FG (Figure 2C)	1	89.8	89.8	73.9	76.4	73.9	76.4
		5	89.5	92.0	72.4	75.6	72.8	75.7
		32	67.1	50.9	31.0	46.3	71.6	71.7
		224	55.4	27.0	15.5	26.5	71.4	72.7

metrics and reconciliation processes are key to consistency in situations like this. Despite these challenges, we are confident that MaintIE represents the most extensive annotated dataset currently available for maintenance texts (Akhbardeh et al., 2020).

8.2. Benchmark Models

The annotated MaintIE corpora was used to facilitate automated IE for engineering maintenance texts and explore the performance of two prominent deep learning architectures for entity and relation extraction: SpERT, a token classification model (Eberts and Ulges, 2020), and REBEL, a sequence-to-sequence model (Cabot and Navigli, 2021). We assessed these models' adaptability to different entity hierarchy sizes and the potential enhancement in fine-grained IE performance when incorporating the coarse-grained corpus for additional fine-tuning. The results, presented in Table 4 and Appendix 10.3, are insightful. As the entity hierarchy expands, there is a reduction in entity recognition efficiency. Yet, the models demonstrated proficiency in relation extraction across the board, as evident from their F1 scores. As expected, untyped models, which classify one or two entity types only, outperformed others, whereas models operating on the complete 224-class entity hierarchy have lower performance. Notably, the sequence-to-sequence model failed for strict relation extraction at the 32 and 224 entity class granular levels. Such performances are likely influenced by the limited corpus size, the diversity of entity classes, and imbalances (details in Table 3), emphasising the challenging nature of this task, even when dealing with short texts.

Our observations suggest that pre-fine-tuning on the coarse-grained corpus (Figure 2C) before transitioning to the fine-grained corpus can enhance performance. Nevertheless, the magnitude of this enhancement varies depending on the metrics applied. For instance, relation extraction performance for the sequence-to-sequence model increased from 0.6 to 51.1 (macro RE Strict) and 70.2 to 72.5 (macro RE Loose) at the 32-class level. The token classification model presented the opposite trend. This difference is likely due to the model architectures and how they learn from data. While REBEL's performance improved with additional training data, which assisted in contextualising the MST from its pretraining corpus, SpERT encountered a decline. This decrease was likely attributed to the fact that during the coarse-grained fine-tuning phase, SpERT established dominant weights for the five top-level classes, making it challenging to re-adjust when subsequently fine-tuned on the fine-grained corpus.

The corpus characteristics, notably the long-tail distribution of entities, present a substantial challenge to the IE algorithms. The detailed distribution of entities in the hold-out test set, as illustrated in Appendix 10.2, accentuates the task's complexity. Despite these challenges, the MaintIE dataset's efficacy in relation extraction from short maintenance texts unveils promising avenues for enhancing technical support and information retrieval for domain experts, along with opportunities for taxonomy extraction and a deeper understanding of asset performance.

While the primary intent of this paper is to introduce the MaintIE scheme and its annotated data rather than to develop models of exceptional performance, we offer these models as prospective benchmarks against which future model performance can be assessed. Future investigations might explore the use of the entity hierarchy to improve entity recognition performance, refine the training strategies to further take advantage of the coarse-grained corpus, or constrain the generation process of the sequence-to-sequence model for optimised entity recognition similar to Josifoski et al. (2022).

9. Conclusion and Future Work

This paper introduced MaintIE, an annotation scheme for extracting information from the low-quality short text field of maintenance work orders. Two annotated corpora were presented: a fine-grained expert-annotated corpus of 1,076 texts comprising 224 entity classes and 6 relations and a coarse-grained large-scale corpus consisting of 7,000 texts comprising 5 classes and 6 relations. The use of these corpora for automated information extraction was explored by training two benchmark deep learning models using token classification and sequence-to-sequence architectures, respectively. The best models achieved over 70% micro and macro F1-scores. Future work seeks to add data and annotations to the MaintIE corpus for the entities not present in the current maintenance short texts by reaching out to other engineering sectors and further exploring training strategies to make use of coarse-grained corpora to improve fine-grained entity and relations-focused information extraction.

10. Acknowledgements

This research is supported by the Australian Research Council through the Centre for Transforming Maintenance through Data Science (grant number IC180100030). Additionally, Bikaun acknowledges funding from the Mineral Research Institute of Western Australia.

Appendices

10.1. Appendix A: Example of Data

This section presents a JSON example from the MaintIE corpus, providing a representative glimpse of the data's structure and format.


```
{
  "tokens": [
    "engine",
    "oil",
    "blender",
    "running",
    "constantly"
  ],
  "entities": [
    {
      "start": 2,
      "end": 3,
      "type": "PhysicalObject/
      MatterProcessingObject/
      MixingObject"
    },
    {
      "start": 0,
      "end": 1,
      "type": "PhysicalObject/
      DrivingObject/
      CombustionEngine"
    },
    {
      "start": 3,
      "end": 5,
      "type": "Process/
      UndesirableProcess"
    },
    {
      "start": 1,
      "end": 3,
      "type": "PhysicalObject/
      MatterProcessingObject/
      MixingObject"
    }
  ],
  "relations": [
    {
      "type": "hasPart",
      "head": 1,
      "tail": 3
    },
    {
      "type": "hasPatient",
      "head": 2,
      "tail": 3
    },
    {
      "type": "isA",
      "head": 3,
      "tail": 0
    }
  ]
}
```


10.2. Appendix B: Distribution of Entity and Relation Annotations

This section illustrates the distribution of entity and relation annotations in the two MaintIE corpora.

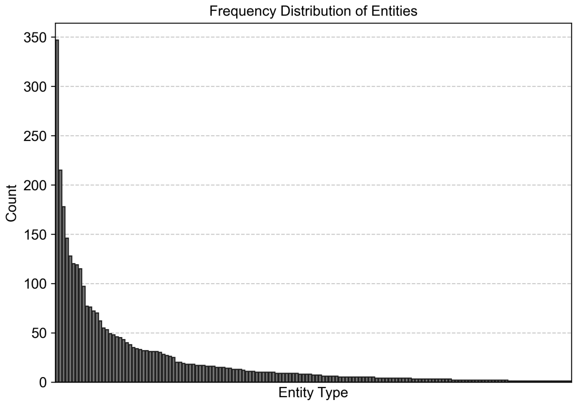

Figure 3: Fine-Grained Expert-Annotated corpus: Distribution of entity annotations.

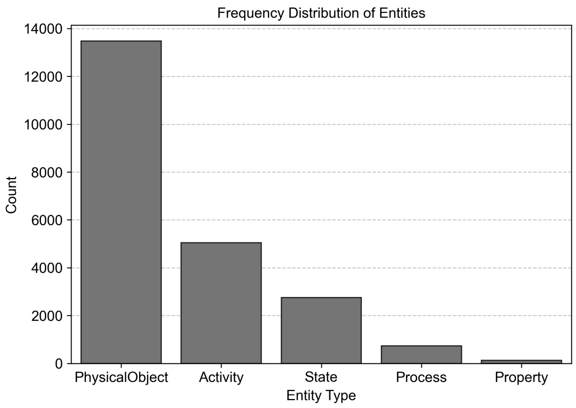

Figure 4: Coarse-Grained Large-Scale corpus: Distribution of entity annotations.

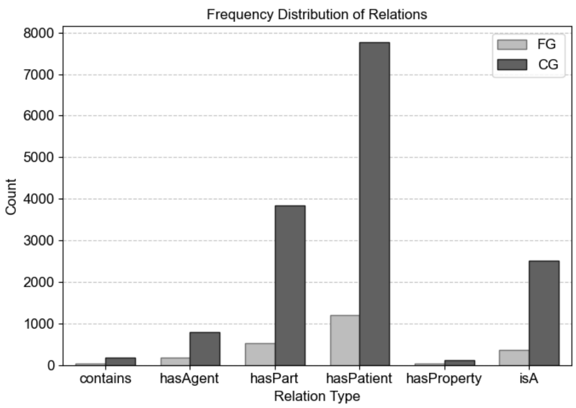

Figure 5: Comparative distribution of relation annotations across Fine-Grained Expert-Annotated and Coarse-Grained Large-Scale corpora.

10.3. Appendix C: Detailed Model Results

This section provides comprehensive results from chosen models to enhance understanding of their performance. Detailed entity recognition results are excluded here due to the extensive range of labels but are fully available in the accompanying repository.

Type	**Prec.**	**Recall**	**F1**	**Support**
contains	0	0	0	3
isA	0	0	0	20
hasPatient	0	0	0	137
hasProperty	0	0	0	5
hasAgent	0	0	0	8
hasPart	0	0	0	61
macro	0	0	0	234
micro	0	0	0	234

Table 5: Test set performance summary of optimal REBEL model (*CG* $\rightarrow$ *FG*) on strict relation extraction: Evaluation across 224 entity classes.

Type	**Prec.**	**Recall**	**F1**	**Support**
contains	100	100	100	3
isA	37	75	49	20
hasPatient	81	85	83	137
hasProperty	80	80	80	5
hasAgent	86	75	80	8
hasPart	75	69	72	61
macro	73	79	76	234
micro	76	81	77	234

Table 6: Test set performance summary of optimal REBEL model (*CG* $\rightarrow$ *FG*) on loose relation extraction: Evaluation across 224 entity classes.

Type	**Prec.**	**Recall**	**F1**	**Support**
contains	100	100	100	3
isA	31	25	28	20
hasPatient	35	21	26	137
hasProperty	0	0	0	5
hasAgent	33	25	29	8
hasPart	34	23	27	61
macro	34	23	27	234
micro	39	32	35	234

Table 7: Test set performance summary of optimal SpERT model (*FG*) on strict relation extraction: Evaluation across 224 entity classes.

11. Bibliographical References

Farhad Akhbardeh, Travis Desell, and Marcos Zampieri. 2020. MaintNet: A collaborative open-source library for predictive maintenance language resources. In *Proceedings of the 28th*

Type	Prec.	Recall	F1	Support
contains	100	100	100	3
isA	50	40	44	20
hasPatient	82	49	61	137
hasProperty	67	80	73	5
hasAgent	50	38	43	8
hasPart	56	38	45	61
macro	70	46	56	234
micro	67	57	61	234

Table 8: Test set performance summary of optimal SpERT model (*FG*) on loose relation extraction: Evaluation across 224 entity classes.

International Conference on Computational Linguistics: System Demonstrations, pages 7–11, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Abhijeet S Bhardwaj, Dharmaraj Veeramani, and Shiyu Zhou. 2023. Confidently extracting hierarchical taxonomy information from unstructured maintenance records of industrial equipment. *International Journal of Production Research*, pages 1–20.

Abhijeet Sandeep Bhardwaj, Akash Deep, Dharmaraj Veeramani, and Shiyu Zhou. 2021. A custom word embedding model for clustering of maintenance records. *IEEE Transactions on Industrial Informatics*, 18(2):816–826.

Tyler Bikaun, Tim French, Melinda Hodkiewicz, Michael Stewart, and Wei Liu. 2021. Lexiclean: An annotation tool for rapid multi-task lexical normalisation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 212–219.

Tyler Bikaun, Melinda Hodkiewicz, and Wei Liu. 2024. MaintNorm: A corpus and benchmark model for lexical normalisation and masking of industrial maintenance short text. In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, pages 68–78, San Ġiljan, Malta. Association for Computational Linguistics.

Tyler Bikaun, Michael Stewart, and Wei Liu. 2022. Quickgraph: A rapid annotation tool for knowledge graph extraction from technical text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Michael P Brundage, Thurston Sexton, Melinda Hodkiewicz, Alden Dima, and Sarah Lukens. 2021. Technical language processing: Unlocking maintenance knowledge. *Manufacturing Letters*, 27:42–46.

Michael P Brundage, Thurston Sexton, Melinda Hodkiewicz, Katherine C Morris, Jorge Arinez, Farhad Ameri, Jun Ni, and Guoxian Xiao. 2019. Where do we start? guidance for technology implementation in maintenance management for manufacturing. In *International Manufacturing Science and Engineering Conference*, volume 58745. American Society of Mechanical Engineers.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381.

Alden Dima, Sarah Lukens, Melinda Hodkiewicz, Thurston Sexton, and Michael P Brundage. 2021. Adapting natural language processing for technical text. *Applied AI Letters*, 2(3):e33.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, volume 2, pages 837–840.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *24th European Conference on Artificial Intelligence: ECAI 2020*, pages 2006–2013. IOS Press.

Yiyang Gao, Caitlin Woods, Wei Liu, Tim French, and Melinda Hodkiewicz. 2020. Pipeline for machine reading of unstructured maintenance work order records. In *Proceedings of the 30th European Safety and Reliability conference*.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Melinda Hodkiewicz and Mark Tien-Wei Ho. 2016. Cleaning historical maintenance work order data for reliability analysis. *Journal of Quality in Maintenance Engineering*, 22:146–463.

IEC. 2019. IEC 81346-2 industrial systems, installations and equipment and industrial products – structuring principles and reference designations – part 2: Classification of objects and codes for classes. Technical report, European Committee for Electrotechnical Standardization, Geneva, Switzerland.

Naresh Iyer, Nurali Virani, Zhaoyuan Yang, and Abhinav Saxena. 2022. Mixed initiative approach for reliable tagging of maintenance records with machine learning. In *Annual Conference of the PHM Society*, volume 14.

Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni, and Robert West. 2022. Genie: Generative information extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4626–4643.

Dan Jurafsky. 2000. *Speech & Language processing*. Pearson Education India.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Paul R Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pages 1989–1993.

Shenae Lee, Maria Vatshaug Ottermo, Stein Hauge, and Mary Ann Lundteigen. 2023. Towards standardized reporting and failure classification of safety equipment: Semi-automated classification of failure data for safety equipment in the operating phase. *Process Safety and Environmental Protection*, 177:1485–1493.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Sarah Lukens, Manjish Naik, Kittipong Saetia, and Xiaohui Hu. 2019. Best practices framework for improving maintenance data quality to enable asset performance analytics. In *Annual Conference of the PHM Society*, volume 11.

Madhusudanan Navinchandran, Michael E Sharp, Michael P Brundage, and Thurston B Sexton. 2021. Discovering critical kpi factors from natural language in maintenance work orders. *Journal of Intelligent Manufacturing*, pages 1–19.

Maria Vatshaug Ottermo, Solfrid Håbrekke, Stein Hauge, and Lars Bodsberg. 2021. Technical language processing for efficient classification of failure events for safety critical equipment. In *PHM Society European Conference*, volume 6, pages 289–297.

Sachin Pawar, Girish K Palshikar, and Pushpak Bhattacharyya. 2017. Relation extraction: A survey. *arXiv preprint arXiv:1712.05191*.

Teresa Robertson and Philip Atkins. 2013. Essential vs. accidental properties. *Stanford Encyclopedia of Philosophy*.

Kittipong Saetia, Sarah Lukens, Erik Pijcke, and Xiaohui Hu. 2019. Data-driven approach to equipment taxonomy classification. In *Annual Conference of the PHM Society*, volume 11, pages 1–13.

Serhad Sarica, Jianxi Luo, and Kristin L Wood. 2020. Technet: Technology semantic network based on patent data. *Expert Systems with Applications*, 142:112995.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin – Madison.

Thurston Sexton, Melinda Hodkiewicz, Michael P Brundage, and Thomas Smoker. 2018. Benchmarking for keyword extraction methodologies in maintenance work orders. *Proceedings of the Annual Conference of the PHM Society*, 10(1).

Christer Stenström, Mustafa Al-Jumaili, and Aditya Parida. 2015. Natural language processing of maintenance records data. *International Journal of COMADEM*, 18(2):33–37.

Michael Stewart, Melinda Hodkiewicz, Wei Liu, and Tim French. 2022. Mwo2kg and echidna: Constructing and exploring knowledge graphs from maintenance data. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, page 1748006X221131128.

Robert B Stone and Kristin L Wood. 1999. Development of a functional basis for design. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 19739, pages 261–275.

Douglas S Thomas. 2018. *The costs and benefits of advanced maintenance in manufacturing*. US Department of Commerce, National Institute of Standards and Technology.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task:

Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Juan Pablo Usuga Cadavid, Bernard Grabot, Samir Lamouri, Robert Pellerin, and Arnaud Fortin. 2022. Valuing free-form text data from maintenance logs through transfer learning with camembert. *Enterprise Information Systems*, 16(6):1790043.

Dario Valcamonico, Piero Baraldi, Enrico Zio, Luca Decarli, Anna Crivellari, and Laura La Rosa. 2024. Combining natural language processing and bayesian networks for the probabilistic estimation of the severity of process safety events in hydrocarbon production assets. *Reliability Engineering & System Safety*, 241:109638.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešic, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoglu, et al. 2021. Multilexnorm: A shared task on multilingual lexical normalization. In *Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509. Association for Computational Linguistics.

C. J. van Rijsbergen. 1979. *Information Retrieval*. Butterworth-Heinemann.

Xiyu Wang and Nora El-Gohary. 2023. Deep learning-based relation extraction and knowledge graph-based representation of construction safety requirements. *Automation in Construction*, 147:104696.

C Woods, M Selway, M Hodkiewicz, F Ameri, and others. 2021. On the notion of maintenance state for industrial assets. In *FOMI 2021: 11th International Workshop on Formal Ontologies meet Industry, JOWO 2021*, pages 1–14.

Caitlin Woods, Matt Selway, Tyler Bikaun, Markus Stumptner, and Melinda Hodkiewicz. 2023. An ontology for maintenance activities and its application to data quality. *Semantic Web Journal*, Pre-press:1–34.

Zhe Yang, Piero Baraldi, and Enrico Zio. 2020. A novel method for maintenance record clustering and its application to a case study of maintenance optimization. *Reliability Engineering & System Safety*, 203:107103.