MaintNorm: A corpus and benchmark model for lexical normalisation and masking of industrial maintenance short text

Tyler Bikaun, Melinda Hodkiewicz, Wei Liu
The University of Western Australia
tyler.bikaun@research.uwa.edu.au

Abstract

Maintenance short texts are invaluable unstructured data sources, serving as a diagnostic and prognostic window into the operational health and status of physical assets. These user-generated texts, created during routine or ad-hoc maintenance activities, offer insights into equipment performance, potential failure points, and maintenance needs. However, the use of information captured in these texts is hindered by inherent challenges: the prevalence of engineering jargon, domain-specific vernacular, random spelling errors without identifiable patterns, and the absence of standard grammatical structures. To transform these texts into accessible and analysable data, we introduce the MaintNorm dataset, the first resource specifically tailored for the lexical normalisation task of maintenance short texts. Comprising 12,000 examples, this dataset enables the efficient processing and interpretation of these texts. We demonstrate the utility of MaintNorm by training a lexical normalisation model as a sequence-to-sequence learning task with two learning objectives, namely, enhancing the quality of the texts and masking segments to obscure sensitive information to anonymise data. Our benchmark model demonstrates a universal error reduction rate of 95.8%. The dataset and benchmark outcomes are made available to the public under the MIT license.[1]

1 Introduction

Industrial user-generated content, such as maintenance work order (MWO) records, logbooks, and incident reports, constitutes a rich repository of data. This data is pivotal for applications in predictive maintenance, safety analysis, process optimisation, and product life cycle management (Brundage et al., 2021). Specifically, in the maintenance sector, MWO short texts (MST) are instrumental in documenting the condition of assets and the maintenance activities performed, as well as informing the design of maintenance strategies. These texts, typically authored by technicians, serve as critical input for future maintenance endeavours. Furthermore, reliability engineers scrutinise these historical records to gain a deeper understanding of equipment failure modes (Lee et al., 2023), enhance root cause analysis (Valcamonico et al., 2024), and develop key performance indicators such as mean-time-to-failure and remaining useful life (Lukens et al., 2019; Bikaun and Hodkiewicz, 2021).

A	AN426 REPLACE BROKEN ALT BOLT **<id> replace broken alternator bolt** R/H Steering Cyl Pin & Brg **right hand steering cylinder pin and bearing**
B	air con belt u/s **air conditioner belt unserviceable** 1000H Mech Insp Carry Roll No 2 RH DN9817 **<num> hour mechanical inspection carry roller number <num> right hand <id>**
C	ZH6907 C/out pos 2 tyre **<id> change out position <num> tyre** Left cab aircon e/leakage flt **left cabin air conditioner electrical leakage fault**

Table 1: User-Generated maintenance short texts for heavy mobile equipment across three companies, with **bold** blue text indicating normalised and masked forms.

Consider Table 1, which showcases examples of MSTs from various companies. These texts, often characterised by technical jargon, domain-specific vernacular, and frequent linguistic inaccuracies, pose significant challenges regarding data quality and processing efficiency (Hodkiewicz and Ho, 2016; Brundage et al., 2021). The resultant ambiguity and lack of standardisation impede effective pattern recognition and trend analysis, impacting maintenance decision-making.

MSTs frequently contain sensitive information, ranging from equipment identifiers to personnel names, raising confidentiality concerns (Brundage

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

et al., 2021). Consequently, there is a scarcity of publicly available industrial (as opposed to governmental) raw MST datasets, with limited examples like MaintNet comprising 7,000 MSTs, and a dataset on excavators with 5,486 MSTs.[2,3] Industrial companies' hesitation to release data, driven by concerns over identification (Sikorska et al., 2020) and a lack of appropriate anonymisation tools, significantly hamper the advancement of technical language models in this critical commercial sector.

Lexical normalisation, the process of transforming non-standard words and phrases into their standard forms (Han and Baldwin, 2011), provides a promising solution for addressing the issue of poor text quality in MSTs. While there has been extensive research on the lexical normalisation of social media texts (Baldwin et al., 2015; van der Goot et al., 2021), industrial maintenance texts have not received similar attention. Currently, state-of-the-art lexical normalisation has been achieved by formulating it as a sequence-to-sequence learning task whereby a sequence with potentially non-canonical (noisy) tokens is transduced into a sequence of canonical (clean) tokens (Samuel and Straka, 2021).

This paper addresses the need to enhance the quality of MST data and simultaneously de-identify sensitive information through a sequence-to-sequence learning approach. Our annotated corpora and model are designed to generate high-quality, normalised sequences with strategically masked segments to obscure sensitive or semantically redundant information. This is particularly crucial in knowledge elicitation for tasks such as information extraction annotation, which rely on domain expertise due to the tacit knowledge needed for interpreting these short texts, whom have limited time resources.

This paper's primary contributions are threefold:

- We introduce the first publicly available annotated corpus for lexical normalisation and masking of maintenance short texts,
- We systematically characterise the lexical noise present in maintenance short texts, and
- We demonstrate the efficacy of sequence-to-sequence language modelling in performing lexical normalisation and masking as a unified task using a structured encoding scheme.

[2] MaintNet Large Technical Database
[3] Prognostics Data Library: Excavator MWOs

2 Background and Related Work

Lexical normalisation, an important task in natural language processing, involves converting non-standard or informal language—such as abbreviations, colloquialisms, and misspellings—into a more standard form. This process is especially pertinent in the context of MSTs, where the prevalence of informal language poses unique challenges (Brundage et al., 2021). Lexical normalisation, as defined by Han and Baldwin (2011), aims to systematically transform non-standard words to their standard equivalents, thereby enhancing readability and facilitating more effective processing for a range of downstream natural language processing applications.

MSTs are key information sources in asset-intensive organisations. Numerous studies, such as those by Hodkiewicz and Ho (2016), Saetia et al. (2019), Gao et al. (2020), and Akhbardeh et al. (2020), have explored the unique lexical challenges these texts present. These works have primarily focused on enhancing MST quality for downstream tasks, employing methods ranging from heuristic approaches to normalisation dictionaries and distance-matching algorithms such as Levenshein (Levenshtein et al., 1966). However, these approaches often lack robustness and adaptability in broader maintenance contexts. Moreover, the confidentiality concerns associated with MSTs remain challenging to address, resulting in a scarcity of publicly available datasets, as highlighted by Akhbardeh et al. (2020) and Brundage et al. (2021).

In contrast to work on MSTs, the field of lexical normalisation has evolved significantly over time. Early insights into the challenges and methodologies were provided by foundational studies like those of Han and Baldwin (2011) and Baldwin et al. (2015). The work of van der Goot et al. (2021) expanded these insights to multilingual normalisation, demonstrating the task's complexity across different languages. The task of lexical normalisation has witnessed a paradigm shift from non-sequence-to-sequence models, such as *MoNoise* by van der Goot (2019), to more sophisticated sequence-to-sequence models (Samuel and Straka, 2021). This transition, highlighted in the work of Lourentzou et al. (2019), marks a critical juncture in the history of lexical normalisation.

The formulation of lexical normalisation as a sequence-to-sequence learning task has led to the use of pre-trained knowledge representations as explored by Muller et al. (2019), and the joint normalisation and masking of e-commerce dialogues by Nguyen and Cavallari (2020). More recently, the Shared Task on Multilingual Lexical Normalization (van der Goot et al., 2021) saw the extensive application of sequence-to-sequence learning predominately through Transformer-based models leveraging pre-trained language models such as (Samuel and Straka, 2021)'s state-of-the-art token-by-token normalisation using ByT5 (Xue et al., 2022) which represents the cutting-edge in the field.

The convergence of these developments in MSTs and lexical normalisation underscores the necessity for adaptable, robust models capable of managing the complexities of maintenance texts. Our research aims to leverage state-of-the-art techniques to improve the lexical quality of MSTs, focusing on joint normalisation and masking to enhance both readability and confidentiality.

3 Data Description

The MaintNorm dataset comprises 12,000 MSTs sourced from three major Australian mining companies.[4] These texts pertain to heavy mobile equipment (HME) – machinery used for operations like excavation, material handling, and earth transportation, including but not limited to haul trucks, dozers, excavators, water trucks, and drill machines. The content of these texts encompasses both routine and ad-hoc maintenance tasks, both planned and executed, as well as insights into the condition of the HME systems and their individual components. Table 1 provides examples sampled from each company.

3.1 Selection

To create the MaintNorm dataset, maintenance texts were randomly selected from a comprehensive repository belonging to the three participating organisations. Each organisation contributed 4,000 texts, ensuring equal representation. The primary objective of this diverse collection is to investigate the feasibility of developing a normalisation and masking model that can effectively operate across different organisational contexts for a given asset type. This approach also helps to discern whether specific models, attuned to the unique linguistic characteristics of each organisation, yield superior results. Detailed corpus statistics, including average text length, vocabulary size, and total token count for each company, are presented in Table 2.

[4]We use A, B, and C to refer to these companies to ensure their privacy.

3.2 Preprocessing

The preprocessing of the MaintNorm corpus was minimal to preserve the raw characteristics of the texts, the texts only underwent basic tokenisation based on whitespace prior to annotation.

3.3 Annotation

Annotation is performed by the first author due to resource constraints. The annotator is experienced with lexical normalisation and industrial maintenance. The annotation tool LexiClean (Bikaun et al., 2021) was used for all lexical normalisation and masking. An overview of the annotated corpora is presented in Table 5. Similar to Han and Baldwin (2011), the following guidelines were used in the annotation process:

Spelling corrections. Canonical forms are adopted to rectify spelling discrepancies within the corpus, such as omissions, redundancies, or incorrect characters. For example, abbreviations like 'eng' are converted to their full form 'engine'.

True casing. The dataset is standardised using true casing, where inappropriate capitalisation is corrected. For instance, 'REPLACE ENGINE' is modified to 'replace engine', except for proper nouns that retain capitalisation, e.g., 'UL123 teleremote' to 'UL123 Tele-Remote'. Acronyms are cased according to their standard usage.

Abbreviation expansion. Maintenance text abbreviations are expanded to their full lexical forms to facilitate uniformity and clarity. For instance, 'c/o' becomes 'change out'.

Concatenation and tokenisation. Incorrectly concatenated multi-word expressions are separated (e.g., 'repair/replace' to 'repair / replace', '250hr' to '250 hour'), enhancing the granularity for downstream tasks such as information extraction.

Token masking. In addition to normalisation, token-level entity masks (tags) were applied to text spans using the scheme in Table 4. The use of token-level entity tags is twofold. First, due to confidentiality concerns, the texts have been preprocessed to obfuscate any identifiers about assets, or-

Company	Length ($\mu \pm \sigma$)	Vocab Size	Tokens	Modified	Norm Only	Mask Only
A	5.2 (1.2)	2,561	20,944	-	-	-
	5.4 (1.3) (↑ 3%)	1,106 (↓ 57%)	21,591 (↑ 3%)	3,998	115	45
B	5.5 (1.4)	3,100	21,919	-	-	-
	6.2 (1.8) (↑ 13%)	1,360 (↓ 56%)	24,690 (↑ 13%)	3,946	192	321
C	5.1 (1.5)	4,168	20,559	-	-	-
	5.5 (1.8) (↑ 7%)	2,048 (↓ 51%)	22,114 (↑ 7%)	3,431	1,879	150
A+B+C	5.3 (1.4)	7,612	63,422	-	-	-
	5.7 (1.7) (↑ 8%)	2,872 (↓ 62%)	68,395 (↑ 8%)	11,375	2,116	586

Table 2: Summary of MaintNorm corpus statistics: This table displays statistics for 4,000 texts from each company, focusing on heavy mobile equipment. It includes token-based text length and vocabulary size. Greyed rows represent post-normalisation and masking statistics. Changes due to normalisation and masking are indicated by arrows and percentages (↑/↓ X%). The right-hand section of the table delineates the text transformations, categorising them as *Modified* for texts undergoing normalisation or masking, *Norm Only* for texts exclusively normalised, and *Mask Only* for texts solely subjected to masking.

N	M	Example
1	1	Single word normalisation, e.g., 'eng' to 'engine'.
1	> 1	Single to multi-word normalisation, e.g., 'c/o' to 'change out'.
1	0	Removal of superfluous characters, e.g., 'T' in 'replace engine T' where 'T' is erroneous.
> 1	1	Concatenation of fragmented words, e.g., 'eng ine' to 'engine'.
> 1	> 1	Combining fragmented words into multi-word normalisations, e.g., 'eng ineoi l' to 'engine oil'.

Table 3: Examples of N:M normalisation transformations in the MaintNorm dataset.

Mask	Description
<id>	Asset identifiers e.g. *ENG001*, *rd1286*
<sensitive>	Sensitive organisation-specific information such as proprietary systems, third-party contractors, names of personnel, etc.
<num>	Numerical digits e.g. *8*, *7001223*
<date>	Numerical and phrase representations of dates e.g. *10/10/2023*, *8th Dec*

Table 4: MaintNorm token masking scheme used for privacy preservation and redundancy removal.

ganisations, personnel, etc, using token-level masking, which was applied in the annotation process. Second, tags such as <num> and <date> reduce the semantic duplication of texts for downstream annotation tasks such as information extraction as maintenance short texts can be generated in very similar fashions such as 'replace pump 1' and 'replace pump 2', here the semantics is the same but there is redundancy when annotating for other tasks. Hence, it is desirable to normalise texts like these to a unified form such as 'replace pump <num>', which represents this structure generally.

3.4 Post-processing and Obfuscation

Two steps were performed post-annotation to ensure the texts were suitable for model training and public release. First, all token-level entity masks were used to mask the respective tokens, e.g. an <id> entity masks on the "PU001" in "replace PU001" would subsequently convert the text into "replace <id>". This process was performed for all masking tokens. Simultaneously, we ensure that masked tokens are obfuscated before public release. We do this by mapping over each text and identifying any masked tokens, which we map to an arbitrary representation of the same semantic type. For example, for <id>, we copy the alphanumerical and cased structure of the original identifier. For <date> and <num>, we copy the structure but permutate it. For <sensitive>, we detect the n-gram size and correspondingly impute a non-sensitive value. These actions ensure that the dataset captures the original essence of the task whilst maintaining a level of desensitisation to allow public release of the dataset.

3.5 Dataset Split

For the purpose of evaluating the generalisation of lexical normalisation and masking within our dataset, we divided it into training, development, and testing sets. Adhering to the conventional split ratio of 80/10/10, our dataset is segmented into 3,200 training texts and 400 texts each for development and testing. Furthermore, we organised the data into distinct company-specific segments (A, B, C) and an aggregated dataset (A+B+C). This segmentation strategy aims to investigate whether the

		A	B	C	A+B+C
Normalisation Operations	Char. addition	3,022	4,704	2,781	10,507
	Char. removal	191	939	247	1,377
	Char. rearrangement	145	118	233	358
	Char. replacement	209	508	231	950
	Token expansion	662	2,264	1,281	4,207
	Token removal	194	97	195	486
	Titled cased	69	118	97	284
	Partial casing added	8	6	9	23
	All casing removed	13,826	9,214	7,098	30,138
	All casing added	4	29	36	69
	No change	1,978	7,187	10,173	19,338
Norm. Transforms	1:1	17,898	12,233	8,694	38,825
	1:N	662	2,264	1,281	4,207
	N:1	194	97	195	486
	N:M	2	4	2	8
	N:0	7	15	6	28
Masking Ops.	<id>	4,055	3,916	1,116	9,087
	<sensitive>	44	25	155	224
	<num>	573	1,349	847	2,769
	<date>	9	2	49	60

Table 5: Summary of the normalisation and masking operations applied to maintenance short texts for each organisation and combined. Tokens can have multiple normalisation operations performed upon them; for example "tlerEMOTE" which is normalised to "Tele-Remote" would have the operations character addition ("tleremote" → "tele-rEMOTE"), all casing removed ("tele-rEMOTE" → "tele-remote") and title casing ("tele-remote" → "Tele-Remote"), representing a 1 : 1 normalisation transformation ("tlerEMOTE" → "Tele-Remote"). *Norm.* and *Ops.* refer to normalisation operations, respectively.

linguistic patterns are consistent across different companies and if such uniformity could enhance the performance of a single, universally-trained model. A positive outcome could encourage industrial entities to collaboratively address this task, yielding mutual advantages.

4 Method

4.1 Task Formulation

In this work, we conceptualise the task of lexical normalisation and masking as an auto-regressive sequence-to-sequence learning task. Our approach involves training a Transformer-based encoder-decoder model to transform potentially noisy input sequences into their normalised counterparts. This methodology is an extension of the approach outlined by De Cao et al. (2020), which employs sentinel brackets for demarcating entity boundaries in auto-regressive entity linking.

We have adapted this approach to suit our specific requirements. Our model defines boundaries around both non-canonical words and phrases, as well as their canonical equivalents. For instance, an input sequence such as 'repl ace eng oil' is normalised to 'replace engine oil'. Using our encoding scheme, the sequence-to-sequence model represents this transformation in its output space as '{ repl ace } [replace] { eng } [engine] oil'. The model's output undergoes post-processing to yield the correctly formatted output, 'replace engine oil', by extracting canonical elements and unchanged tokens, as shown by '{ repl ace } [**replace**] { eng } [**engine**] **oil**'. This encoding technique and its application to various normalisation transformations is exemplified in Table 6.

Operation	
1:1	`{ reply } [ replace ]`
N:1	`{ repl ace } [ replace ]`
1:M	`{ repleng } [ replace engine ]`
N:0	`{ $$ } [ ]`
N:M	`{ rep&re pl } [ repair and replace ]`

Table 6: Examples of the normalisation encoding scheme applied to different normalisation operations. Curly brackets ({}) denote a non-canonical span, whereas square brackets ([]) denote a canonical span.

While directly translating into normalised sequences (e.g., 'repl ace eng oil' → 'replace engine oil') may seem straightforward, it poses challenges for evaluation (see Appendix B). Ensuring alignment between input and output sequence translations is a complex task, as highlighted in the work of Sabet et al. (2020). Our encoding scheme directly addresses this challenge by explicitly cap-

turing these transformations. Furthermore, our approach is particularly effective in token masking, as it naturally extends to an N:M operation (e.g., '{ UD01 } [<id>]', '{ blwnEN1 } [blown <id>]').

This methodology contrasts with the token-by-token normalisation strategy of Samuel and Straka (2021). Our approach requires only a single pass through the model, with the output sequence autoregressively generated via beam search decoding. Using this approach, each normalisation is conditioned on one another through the context provided by preceding tokens. This means that the model considers the entire input sequence and the part of the output sequence it has generated to predict each subsequent token. This contextual awareness allows for more cohesive and contextually appropriate normalisations, as the model can use the broader context to resolve ambiguities and infer the most probable normalisation for each token. In contrast, a token-by-token approach normalises each token in isolation, potentially missing the nuances of wider textual context.

4.2 Prefix Constrained Decoding

Building on the framework established by De Cao et al. (2020), our study also explores the use of prefix-constrained decoding to curtail the potential for model hallucination and ensure the alignment of input and output sequences. Prefix-constrained decoding is a technique where text generation is guided by constraints such as prefix tries or heuristics to ensure generated output adheres to specific conditions. This technique can be applied to maintain the alignment of input and output sequences during the decoding process for lexical normalisation. In contrast to entity linking, which relies on a closed set of semantic types to constrain generation, we experiment with this technique to limit the model to uncontrolled generation when generating a normalisation or masking pair; otherwise, it must copy the input sequence verbatim. The efficacy of prefix constraints in enhancing linguistic tasks, including entity recognition (Josifoski et al., 2022) and semantic parsing (Scholak et al., 2021), has been well-documented, supporting their application in our study.

4.3 Model Implementation and Parameters

We implement our sequence-to-sequence model as a Transformer encoder-decoder using the pre-trained foundational model of ByT5 (Xue et al., 2022). ByT5, a token-free model, operates directly on byte sequences, enhancing its capacity to handle various languages and character sets without tokenization. All experiments and models are implemented using PyTorch and the Transformers library (Wolf et al., 2020) using PyTorch Lightning (Falcon, 2019) executed on a single Nvidia GeForce RTX 4080 graphics card. We use *google/byt5-small*, containing 299M parameters, fine-tuned in batches of 16 sequences.[5] Model optimisation uses AdamW with cross-entropy loss and a linear learning rate scheduler. Both source and target sequence lengths are set to 256 tokens, and the model runs for 20,000 steps with early stopping based on validation loss, employing a patience of 5 epochs. Our experiments with prefix constraints use logit renormalisation.

4.4 Evaluation

To measure the generalisation ability of a sequence-to-sequence model trained on our corpus, we evaluate them on the intrinsic word-level error reduction rate (E.R.R.), precision, and recall (van der Goot, 2019).[6] Here, E.R.R. is formulated as:

$$E.R.R. = \frac{TP - FP}{TP + FN} \quad (1)$$

E.R.R. values span from -1 to 1, with negative values indicating predominant incorrect normalisations by the model. A zero score signifies no alterations made by the model, and a score of 1 denotes perfect normalisation. In practice, we use the script provided as part of the Multilingual Shared Task (van der Goot et al., 2021), where we translate the encoded sequences into the traditional newline and tab-separated normalisation format for evaluation.

4.5 Baselines

To evaluate the performance of our sequence-to-sequence model on the MaintNorm corpus, we compare it against three normalisation methods:

Leave-As-Is (LAI): The LAI technique is characterised by its direct approach, retaining the original input without modification, resulting in a nominal E.R.R. of 0%.

Most Frequent Replacement (MFR): MFR employs a lexical database that associates each unigram (individual word) in the input with its most commonly observed replacement in the training

[5] HuggingFace google/byt5-small

[6] See Appendix A for evaluation details.

Company	Extra Data	MaintNorm (ours)			LAI			MFR			ÚFAL		
		P	*R*	*E.R.R.*	*P*	*R*	*E.R.R.*	*P*	*R*	*E.R.R.*	*P*	*R*	*E.R.R.*
A	N	**99.9**	95.8	95.2	0	0	0	**99.9**	91.7	90.9	99.8	92.0	91.0
	Y	**99.9**	**98.1**	**96.6**	0	0	0	**99.9**	91.7	90.9	99.9	92.1	91.3
B	N	98.9	94.6	90.0	0	0	0	**99.8**	93.9	90.2	99.6	91.0	85.5
	Y	99.7	**98.1**	**96.6**	0	0	0	**99.8**	93.9	90.2	99.6	91.7	86.5
C	N	99.4	95.2	89.1	0	0	0	**99.5**	89.9	78.6	99.4	86.6	71.9
	Y	**99.5**	**96.8**	**92.4**	0	0	0	**99.5**	89.9	78.6	99.1	86.6	71.5
A+B+C	-	99.7	97.5	95.8	0	0	0	99.8	93.0	89.4	99.5	90.2	85.0

Table 7: Summary of experiments evaluated on the respective hold-out test sets. *Extra data* refers to using the combined training data (A+B+C) but evaluated on the specific portions test-set. P, R, and E.R.R. refer to the precision, recall, and error reduction rate, respectively. **Bold** denotes the best-performing metric for each company.

corpus. During operation, the system substitutes each word with its prevalent counterpart. When an input word is novel and lacks a precedent in the database, it remains unaltered.

ÚFAL: The ÚFAL model (Samuel and Straka, 2021), based on the ByT5 pretrained language model (Xue et al., 2022), employs a token-by-token normalisation approach. It normalises each word separately, encapsulating it within specific tags for processing by ByT5, aligning with ByT5's pre-training objectives. Recognised as a leading method for multilingual lexical normalisation (van der Goot et al., 2021), ÚFAL was fine-tuned for our experiments using its default settings but without implementing its data augmentation strategies, which we reserve for future exploration.

5 Results

In this section, we examine the outcomes derived from developing the MaintNorm annotated corpus and our implementation of sequence-to-sequence modelling for lexical normalisation and masking within MSTs. The central objectives of this analysis are to address two key questions: Firstly, *what are the defining characteristics of lexical noise present in MSTs?* Secondly, *how effective is the application of sequence-to-sequence language modelling in executing lexical normalisation and masking as a combined task?*

5.1 MaintNorm Corpus Construction and Characterisation

In constructing the MaintNorm corpus, a significant observation across all three participating companies was their non-standard approach to casing. As highlighted in Table 5, the most prevalent normalisation operation involved the complete removal of casing, indicative of an excessive use of capital letters. It is noteworthy, however, that while fully capitalised tokens are rare within the corpus, they do occur and typically denote domain-specific acronyms such as 'TECO' (technically completed) and 'HAZ' (hazard), which are essential for domain experts.

Regarding the nature of normalisation transformations, MaintNorm primarily exhibits minimal N:M transformations, mirroring the tendencies observed in the WNUT corpus (Baldwin et al., 2015). This trend suggests a predominance of simpler, more direct normalisation methods within the corpus. Notably, a significant portion of the texts in MaintNorm, accounting for 94.8%, underwent at least one normalisation or masking operation. This rate was particularly high in two companies (A and B), where almost all texts in their respective portions of the corpus were subject to these operations. This extensive normalisation and masking process led to a substantial reduction (>50%) in vocabulary size across all three companies. This reduction underscores the impact of normalisation and masking on the diversity and complexity of the corpus vocabulary.

Table 5 further reveals that the characteristics of noise and masking in the maintenance communication language are consistent across companies despite their independent creation. The distributions of normalisation and masking operations highlight similar characteristics, such as the prevalence of normalisation through 1:1 transformations with high frequencies of character additions, whilst also having a high proportion of masks in the forms of `<id>`. Although `<sensitive>` and `<date>` masks appeared less frequently than `<id>` and `<num>`, their inclusion is crucial for maintaining privacy.

5.2 Sequence-to-Sequence Modelling

Here, we discuss the aspects of generalisation for a sequence-to-sequence model on the MaintNorm corpus. An overview of the experimental results is outlined in Table 7.

Comparative analysis with baseline methods. The sequence-to-sequence language model showcased notable efficiency in unified lexical normalisation and masking, achieving an E.R.R. above 90% across all experiments (refer to Table 7). Although the MFR baseline displayed unexpectedly robust performance, the difference between it and our model highlights the presence of non-mappable tokens. This suggests that the task of normalisation and masking may not be exceedingly challenging, which, albeit potentially less stimulating for researchers, is encouraging for practitioners aiming to implement these findings.

In contrast to our approach, MFR, akin to methods in prior studies (Hodkiewicz and Ho, 2016; Saetia et al., 2019; Akhbardeh et al., 2020), relies on dictionary replacement and cannot adapt to dynamic contexts with variable vocabularies. Using the same foundational model as the ÚFAL model allows for directly comparing encoding schemes. The results in Table 7 indicate superior performance of our encoding scheme across all dataset segments, likely due to its ability to contextually process the entire sequence during decoding, unlike ÚFAL's token-by-token method.

Although our model and encoding scheme are effective, we anticipate further enhancements by increasing the size of the pretrained language model and the number of beams in beam search decoding, which was limited to three due to resource constraints.

	A	B	C
Incorrect Predictions	56/2,059	48/2,202	70/2,050
Normalisation Errors	50	46	53
Masking Errors	6	2	7

Table 8: Error analysis of the best-performing models on their respective test sets from Table 7.

Comparative analysis: individual vs combined models. Evaluating model performance for individual companies against a unified model reveals distinct advantages in adopting a single, combined approach. This consolidated model notably enhances normalisation and masking capabilities, evidenced by a 1.4-6.6 E.R.R. improvement when leveraging additional training data. Although the single model approach appears superior, the performance of organisation-specific models, which closely rival the combined model using only a third of the data, is also noteworthy. Identifying the exact contributors to these performance disparities is challenging. However, qualitatively examining the corpora indicates common language use across the companies. This linguistic similarity suggests that merging the datasets creates a more substantial and varied corpus, enhancing the model's ability to generalise effectively.

Analysis of model errors. Despite achieving high precision and recall in normalisation and masking (see Table 7), our models are not entirely error-free, with error rates ranging from 2.2% to 3.4%. Table 8 outlines these errors. A closer qualitative analysis of incorrect predictions revealed that many errors originate from hapaxes and hapax legomena, causing inaccuracies or missed normalisation and masking opportunities. A common error pattern involves incorrectly handling concatenated corrections (e.g., '&8on' → 'and <num> on', '80A' → '<num> amperage'). Enhancing the MaintNorm corpus with a more diverse range of text samples will likely improve model performance by introducing a wider variety of linguistic scenarios, reducing the potential for such errors.

Effectiveness of encoding scheme and prefix-constrained decoding. Our model's high performance on the MaintNorm corpus, using a specific encoding scheme for lexical normalisation and masking, demonstrates its effectiveness in an autoregressive sequence-to-sequence framework. Although a notable challenge arises in data-scarce scenarios, the model struggles with encoding scheme assimilation, necessitating prefix-constrained decoding (see Table 9 in Appendix C). This issue could be mitigated through techniques such as pre-fine-tuning the models on synthetically generated corpora, following approaches similar to Dekker and van der Goot (2020) and Samuel and Straka (2021), and curriculum learning (Bengio et al., 2009). Our main experiments, as detailed in Table 7, achieve optimal results without constraints, benefiting from a robust training dataset.

While prefix-constrained decoding can effectively prevent hallucination and deviations from the encoding scheme, thereby avoiding misalignments between input and output sequences, its im-

plementation is challenging. One notable issue is the degradation in error reduction efficiency, likely caused by logit renormalisation over constrained tokens. Our findings suggest that while the encoding scheme is effective for larger datasets of short texts, its application to smaller or complex corpora warrants further research. Although untested on other normalisation corpora like those in the multilingual shared task (van der Goot et al., 2021), we believe in the scheme's potential adaptability and plan to explore this in future work.

6 Conclusion and Future Work

In this paper, we have introduced the first corpus for normalising and masking maintenance short texts (MST), comprising 12,000 texts from the Australian mining and mineral processing sector. Our findings show that a unified approach to lexical normalisation and masking, using an encoder-decoder Transformer-based language model, delivers high performance on MSTs, surpassing existing state-of-the-art on our custom-constructed corpus. This methodology offers a viable pathway for industrial organisations to manage risk while releasing data, thereby facilitating research on technical language models in this vital commercial sector. We have made our code, corpus, and models publicly accessible under the MIT license. Looking ahead, we envisage expanding the scope of this dataset to encompass diverse maintenance contexts and enriching it with annotations from a broader range of annotators, which we believe will further augment its utility.

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. Bikaun and Liu acknowledge the support from ARC Discovery Grant DP150102405.

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

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

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 and Melinda Hodkiewicz. 2021. Semi-automated estimation of reliability measures from maintenance work order records. In *PHM Society European Conference*, volume 6, pages 9–9.

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

N De Cao, G Izacard, S Riedel, and F Petroni. 2020. Autoregressive entity retrieval. In *ICLR 2021-9th International Conference on Learning Representations*, volume 2021. ICLR.

Kelly Dekker and Rob van der Goot. 2020. Synthetic data for english lexical normalization: How close can we get to manually annotated data? In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6300–6309.

William A Falcon. 2019. Pytorch lightning: The lightweight pytorch wrapper for high-performance ai research.

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 and 15th. Probabilistic Safety Assessment and Management Conference*. ESRA PSAM.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual meeting of the Association for Computational Linguistics: Human language technologies*, pages 368–378.

Melinda Hodkiewicz and Mark Tien-Wei Ho. 2016. Cleaning historical maintenance work order data for reliability analysis. *Journal of Quality in Maintenance Engineering*, 22(2):146–163.

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.

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.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Ismini Lourentzou, Kabir Manghnani, and ChengXiang Zhai. 2019. Adapting sequence to sequence models for text normalization in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 335–345.

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.

Benjamin Muller, Benoît Sagot, and Djamé Seddah. 2019. Enhancing bert for lexical normalization. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 297–306.

Hoang Nguyen and Sandro Cavallari. 2020. Neural multi-task text normalization and sanitization with pointer-generator. In *Proceedings of the First Workshop on Natural Language Interfaces*, pages 37–47.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *EMNLP 2020*, pages 1627–1643.

Kittipong Saetia, Sarah Lukens, Erik Pijcke, and Xiaohui Hu. 2019. Data-driven approach to equipment taxonomy classification. In *Proceedings of the PHM Society Conference*.

David Samuel and Milan Straka. 2021. Úfal at multilexnorm 2021: Improving multilingual lexical normalization by fine-tuning byt5. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 483–492.

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. Picard: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9895–9901.

Joanna Sikorska, Sam Bradley, Melinda Hodkiewicz, and Ryan Fraser. 2020. Drat: Data risk assessment tool for university–industry collaborations. *Data-Centric Engineering*, 1:e17.

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. 2019. Monoise: A multi-lingual and easy-to-use lexical normalization tool. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 201–206.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

A Description of Evaluation Metrics

To assess the effectiveness of our models, we used precision (*P*), recall (*R*), and error reduction rate (*E.R.R.*), following the methodology outlined in (van der Goot, 2019). These metrics offer a comprehensive evaluation of test accuracy. Precision measures the accuracy of the normalisation model's replacements, while recall determines the model's ability to identify and correctly normalise anomalies. Together, these metrics complement the E.R.R., addressing its limitations in distinguishing between over-normalisation and under-normalisation. The definitions of precision, recall, and error reduction rate are as follows:

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

$$E.R.R. = \frac{TP - FP}{TP + FN} \tag{4}$$

Here, the TP (True Positive), FP (False Positive), and FN (False Negative) values are evaluated at the token level. They are conceptualised as follows:

- True Positive (TP): Words that required normalisation and were accurately normalised by the model.
- False Positive (FP): Words incorrectly normalised by the model despite not requiring normalisation.
- False Negative (FN): Words that required normalisation but were either inaccurately normalised or overlooked by the model.

B Description of Alignment Errors

Alignment errors arise when there's a mismatch between the input portion of the model's prediction and the ground truth, posing challenges to accurate evaluation. These errors can occur even when the model's normalisation predictions are technically correct, leading to complexities in the assessment process. The following examples demonstrate how alignment errors manifest within our encoding scheme:

1. Input: "`repl eng oil`"
2. Output (aligned, ground truth): "`{ repl } [ replace ] { eng } [ engine ] oil`"
3. Output (aligned, incorrect): "`{ repl } [ replacement ] eng oil`"
4. Output (misaligned, correct): "`replace { eng } [ engine ] oil`"
5. Output (misaligned, incorrect): "`rep { eng } [ engine ] oil`"

In these cases, converting the encoded outputs to the normalisation format of the shared task (van der Goot et al., 2021) results in alignment issues. For instance, example (4) shows a misalignment where the ground truth aligns "repl" to "replace", but the misaligned output aligns "replace" to "replace". As a result, such instances are incompatible with the evaluation script used in the shared task.

C Analysis of Alignment Errors

In Table 9, we analyse the correlation between the size of the corpus and alignment errors in our model. It's clear that a sufficiently large corpus enhances the model's comprehension of the encoding scheme, reducing alignment errors. This is primarily due to the model's improved ability to avert hallucination and the creation of incorrect structures in normalisation. On the other hand, with smaller corpora, the model is more prone to alignment errors. To counter this in smaller datasets, we implement prefix constraints in our encoding scheme. This method steers the model towards more precise alignment, thereby ensuring output accuracy even with limited data.

However, our analysis also reveals that while prefix-constrained decoding is beneficial for alignment, it may affect the model's overall error-reduction capabilities. This relationship between alignment accuracy and error reduction under prefix constraints poses an interesting area for future research.

Train Fraction	Train Size	Alignment Errors
0.1	960	179/1,200 (14.9%)
0.2	1,920	67/1,200 (5.6%)
0.3	2,880	56/1,200 (4.7%)
0.4	3,840	10/1,200 (0.8%)
0.5	4,800	7/1,200 (0.5%)
0.6	5,760	11/1,200 (0.9%)
0.7	6,720	4/1,200 (0.3%)
0.8	7,680	4/1,200 (0.3%)
0.9	8,640	0/1,200 (0.3%)
1.0	9,600	0/1,200 (0.0%)

Table 9: Overview of alignment errors in relation to corpus size, using a model trained on the combined corpus (A+B+C) and tested with a beam size of 3.