Automatic normalization of noisy technical reports with an LLM: What effects on a downstream task?

Mariame Maarouf[1,2,3] and **Ludovic Tanguy**[1]
[1] CLLE: CNRS & University of Toulouse, France
[2] Centre National d'Études Spatiales (CNES)
[3] MeetSYS
mariame.maarouf@univ-tlse2.fr
ludovic.tanguy@univ-tlse2.fr

Abstract

This study explores the automatic normalization of noisy and highly technical anomaly reports by an LLM. Different prompts are tested to instruct the LLM to clean the text without changing the structure, vocabulary or specialized lexicon. The evaluation of this task is made in two steps. First, the Character Error Rate (CER) is calculated to assess the changes made compared to a gold standard on a small sample. Second, an automatic sequence labeling task is performed on the original and on the corrected datasets with a transformer-based classifier. If some configurations of LLM and prompts can reach satisfying CER scores, the sequence labeling task shows that the normalization has a small negative impact on performance.

1 Introduction

This study focuses on the automatic cleaning of technical and noisy texts and its impact on an automatic fine-grained semantic labeling task. The goal is to assess the capacity of a generative LLM to automatically rectify noise phenomenons in technical texts that are going to be automatically processed afterwards.

The dataset used in this work is composed of French written anomaly reports during Ariane 5 rocket maintenance operations. These types of maintenance records have proven to be not only filled with words from a technical specialized lexicon, but also extremely noisy (text in uppercase, missing accents, spelling errors and misuse of punctuation) (Bikaun et al., 2024b). As such, we chose to explore the cleaning of the noise by an automatic rectification task performed by prompting a generic pretrained large language model (LLM). Three different LLMs were evaluated, with four different prompts covering different levels of information. A first intrinsic evaluation compared the output of the LLM compared to a gold standard. A second, extrinsic evaluation consisted in measuring the performance of an automatic sequence labeling task. We compare the same classifier when trained and applied to the corrected versus original of the annotated data. The results of the automatic semantic labeling allow us to determine whether these corrections are beneficial to the fine-grain semantic analysis of those texts.

This article is organized as follows. Section 2 is a short review of related work on noise in technical text data. In section 3, we present the dataset of French anomaly reports and the methods used to correct the noise. The results of the intrinsic evaluation are presented in Section 4 and the sequence labelling task and its results are presented in Section 5.

2 Related work

Reporting anomalies is a common procedure in the space and aviation domain, as it is encouraged and has become part of the general culture among professionals. Numerous studies have been conducted on using NLP (Natural Language Processing) on aviation anomaly reports showing that a number of different techniques can be of use in the treatment of such texts (Yang and Huang, 2023), ranging from text classification to information retrieval (Tanguy et al., 2016), (Persing and Ng, 2009). The same kind of anomaly reports dataset used in this work, focusing on maintenance operations on Ariane 5 rockets, has already been the object of NLP experiments in Kurela et al. (2020); Galand et al. (2018) but with other objectives (assessing risk level) and based on a coarser grain text analysis. Maintenance reports have also been shown to be particularly noisy and technical texts (Bikaun et al., 2024b) (Akhbardeh et al., 2020), and thus are difficult to process by the usual NLP pipelines conceived for (and from) standardized texts (Brundage et al., 2021), (Dima et al., 2021).

Several studies have already explored ways to clean this type of texts, from rule-based approaches (Hodkiewicz and Ho, 2016) to lexical normalization techniques (Bikaun et al., 2024a). The use of generative LLMs for correction has also been studied, for post-OCR noisy texts in Thomas et al. (2024) and Zhang et al. (2024) and seems to provide better error reduction rates. Bolding et al. (2023) has also shown promising results in the use of an LLM to clean noisy texts while preserving their semantic integrity. Wang et al. (2024) confirm these results, but also shows that the performance of the LLM varies according to the type of noise and that some models have a better ability to perform this task.

3 Corpus, noise and automatic normalization

Our dataset contains 1050 anomaly reports written in French in an industrial setting. This sample was randomly extracted from a much larger database with tens of thousands similar items. These reports are produced systematically every time an irregularity (however trivial) is encountered by an operator in a critical environment. As can be seen in Table 1, a report consists of a short description of a problem (average length = 19.3 words per report) filled with acronyms, components identifiers and specialized lexicon, mostly in telegraphic-like speech, as can be expected in a workplace communication between professionals (Falzon, 1987). But different noise phenomenons are also commonly found: the text is mostly in uppercase, accents are absent, punctuation and spacing is not respected, and some spelling errors can be found. These phenomenons can be explained by a number of factors related to the conditions in which these texts are typed and formatted. The goal of this study is to test if normalizing the text without reformulating or changing the meaning of the text is possible and beneficial to its analysis. The usual preprocessing techniques have proven to be of limited efficiency on this kind of texts, and run the risk of losing too much information (Brundage et al., 2021). For example, an attempt at POS-tagging on our dataset with *Stanza* (Qi et al., 2020) resulted in a 20% error rate.

For this experiment, we selected three small-sized quantized LLMs that could be run locally on a workstation (a constraint due to the confidentiality of the target data), able to process French and which reach state-of-the-art performance in generic benchmarks: *Meta-Llama-3.1-8B-Instruct*, *Meta-Llama-3-8B-Instruct* (Dubey et al., 2024) and *Mistral-7B-Instruct-v0.3* (Jiang et al., 2023). For each model, four different prompts were defined with incrementally additional information. The first one included only the context of creation of this dataset (i.e. operators reporting anomalies during the maintenance of a rocket) and the requested task (i.e. remove the noise phenomenons without altering the meaning). In the second prompt we added the goal of this operation : to prepare the text for further processing by a non-specified NLP program. For the third one, a list of the different expected types of noise to rectify was given. And finally, in the fourth prompt, two reports and their rectified versions were given as examples (few-shot prompting) (*cf.* Appendix A). As is commonly recommended in such cases, all four prompts were written in standard English, with the explicit indication that the source and target texts are in French (Jin et al., 2024). LLM temperature was set to zero, resulting in deterministic outputs and thus not requiring several runs.

4 Intrinsic evaluation of the correction

For a first evaluation of the results, the Character Error Rate (CER) was calculated on a gold standard of 15 manually normalized reports by the authors. A selection of reports were chosen randomly and 15 were selected to get a representation of all the different noise phenomenons. The correction process is not trivial as each word needs to be corrected, at least by putting the correct case back, decisions have to be made regarding abbreviations and punctuation, some words can be ambiguous due to the lack of accents... Table 2 gives the average CER score for the original and each prompting of the LLMs. The scores of the three different LLMs are close, the variations rely on the prompts. As expected, the first two prompts (with less developed instructions) give high error rates, close to the original reports score, which shows a lot of differences with the gold standard. The indication of a post-processing goal in the second prompt did not improve the results and even seems to have worsened it. However, the addition of the list of phenomenons to consider lead to significant improvement. The insertion of examples in the prompt was not efficient though, and even costed a few points to the results, except for Mistral.

Description of the anomaly
PONT 150 KN :DEFAUTS D'ISOLEMENT SUR LES MOTEURS SUIVANTS :MO12 : DIRECTION GV ===>2,8MohmMO13/14 : DIRECTION MV/PV ===>2,6Mohm ET 1,5MohmMO9 : LEVAGE GV ===>4,7MohmMO6/8 : TRANSLATION MV/PV ===> 4,5Mohm ET 2,3MohmNORME : ISOLEMENT MINI > 5 Mohm
DEGRADATION BETON DESSUS CARNEAUX :1) DESSUS CARNEAU EAP2.JPG2) DESSUS CARNEAU NORD 1.JPG AFFAISEMENT GENERAL3) DESSUS CARNEAU NORD.JPG

Table 1: Examples of anomaly reports

Dataset	Character Error Rate
Original Reports	0.43
Llama 3 prompt 1	0.32
Llama 3 prompt 2	0.35
Llama 3 prompt 3	**0.06**
Llama 3 prompt 4	**0.08**
Llama 3.1 prompt 1	0.35
Llama 3.1 prompt 2	0.34
Llama 3.1 prompt 3	0.10
Llama 3.1 prompt 4	**0.08**
Mistral prompt 1	0.28
Mistral prompt 2	0.39
Mistral prompt 3	**0.07**
Mistral prompt 4	**0.07**

Table 2: CER of the automatic rectification

Dataset	Output	CER
Original report	SYNTHESE HSY062 A OFF ATTENDU A ON.	0.35
Gold report	Synthèse HSY062 à OFF attendue à ON.	
Llama3 prompt 1	Synthèse HSY062 : À l'occasion de l'arrêt attendu à bord.	0.47
Llama3 prompt 2	Synthèse HSY062 : Analyse d'anomalie - À l'occasion de l'offre attendue à ce moment-là.	0.64
Llama3 prompt 3	Synthèse HSy062 : À off attendu à on.	0.20
Llama3 prompt 4	Synthèse HSY062 à off attendu à on.	0.13

Table 3: Examples of correction

The examples in Table 3 show one report and its rectification proposed by Llama 3 according to the different prompts. The text to correct was "SYNTHESE HSY062 A OFF ATTENDU A ON." (tr. "Synthesis HSYP62 on OFF expected on ON"). In this particular report, the expected corrections were limited: put back the lowercase and put back the accents on "Synthèse" and the two occurrences of "à". As already stated, the two first prompts produced less accurate corrections. In this case (which is an extreme one) the output contains additional words and substantial changes in meaning (tr. "on the occasion of the stop expected on board" for prompt 1 and "Anomaly analysis - On the occasion of the offer expected at this moment" for prompt 2). This behavior may even be considered as an hallucination. Their respective CER scores are 0.47 for prompt 1 and 0.64 for prompt 2. Prompts 3 and 4 got almost perfect results, although they respectively obtained 0.2 and 0.13 CER. In prompt 3, the punctuation ":" was added, which could be considered acceptable, and one of the acronym letter was put in lowercase. For both of the prompts, "on" and "off" were put in lowercase, which is not the case for the gold but can hardly be considered a mistake. This first intrinsic evaluation allowed us to identify a subset of promising configurations: we arbitrarily consider for the extrinsic evaluation the 5 which obtained a CER of less than 0.1 (indicated in boldface in Table 2).

5 Evaluation on a downstream task

The second experiment conducted in this study consists in an automatic annotation of the 6 datasets (the original reports and the five datasets with a low CER) through a sequence labeling task. The original reports dataset was manually annotated based on a twelve-class typology of sequences. These classes are related to the main type of technical problem reported (i.e. "leakage", "malfunction", "missing component"...). The annotated text segments are lexical markers (cues) of the class ("leak", "leaking", "absence", "missing", "not present"...). The annotation was performed by three linguists. The inter-annotator agreement between the linguists and a field expert was measured with a *gamma* score (Mathet et al., 2015) of 0.63. In the first example in Table 1, the trigger is "DEFAUT" (tr. "DEFECT") and in the second one, "DEGRADATION". Over the 1050 reports, a total of 1406 segments were identified (1114 are used for training, 292 for testing, with an unbalanced distribution of categories). Several fine-tuned transformer-based token classifiers[1] were tested for this task on the original reports dataset with no preprocessing other than folding the whole text in lowercase. The two models that gave the best results for the original corpus on a token-level evaluation were *bert-base-multilingual-uncased* (Devlin et al., 2019) and *camembert-large* (Martin et al., 2020)

[1]Hyper-parameters: learning-rate = $1e-5$, epoch = 20

Classifier	bert-base-multilingual-uncased			camembert-large		
Dataset	Precision	Recall	F-score	Precision	Recall	F-score
Original reports	0.72	0.77	**0.74**	0.71	0.79	**0.74**
Llama 3 prompt 3 (list)	0.60	0.67	0.64	0.66	0.78	**0.71**
Llama 3 prompt 4 (examples)	0.63	0.73	**0.68**	0.58	0.74	0.65
Llama 3.1 prompt 4 (examples)	0.58	0.66	0.62	0.64	0.77	**0.70**
Mistral prompt 3 (list)	0.57	0.66	0.62	0.63	0.73	**0.68**
Mistral prompt 4 (examples)	0.62	0.67	0.64	0.62	0.74	**0.68**

Table 4: Sequence labeling classifier scores

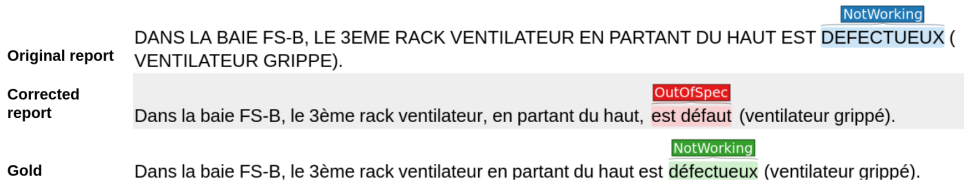

Figure 1: Example of automatic correction impacting the annotation

with respectively 0.68 and 0.67 macro-average F1. However, given the nature of the manual annotation task, the precise segment boundaries may vary without meaningful differences. As such, to get a more accurate view of the scores, the "nervaluate" metric was used, and especially the "entity-type" measure[2] to compute a labeled sequence based evaluation. This measure considers that a sequence which overlap the gold data is a true positive if the type (class) is correct. For the two selected models *bert-base-multilingual-uncased* and *camembert-large*, the entity-type macro-average F1 are both 0.74 (Table 4). Given the tight results, we selected these two classifier configurations to perform the automatic labeling on the corrected datasets.

To reverberate the manual annotation from the original reports to the corrected reports, adjustments had to be done. To that effect, the corresponding offsets of the target sequence in the corrected versions were determined based on the output of the GNU wdiff utility[3], with local homothety transformations for adapting to insertions and deletions. The sequence labeling task was then re-evaluated, using the five corrected versions of both the training and test sets, without any other preprocessing except using the lowercase for *bert-base-multilingual-uncased*.

The results of the automatic sequence labeling shown in Table 4 indicate that the rectification did not increase the scores (best F1 scores in boldface). Instead, they show slightly lower scores for the best two configurations, with 0.71 and 0.70 F1 against 0.74 attained with the original reports dataset. The *camembert-large* classifier obtains overall better results on the rectified data. Potential reasons to this would be that the model has been trained specifically for French language and as such is able to handle the accents, whereas *bert-base-multilingual-uncased*'s tokenizer strips them. Moreover, this BERT model is uncased, which was not an issue for the original reports that were all in uppercase, but for the rectified datasets, the case issues have been corrected. As such, *camembert-large* benefits from more precise and less ambiguous formulations. The decrease of the overall scores also implies a loss of semantic information during the normalization process that impacts the performance of the labeling task.

6 Conclusion

In this study, we have demonstrated that the automatic correction of a technical and noisy text with an LLM produces mitigated results. The scores given by the CER seemed satisfactory enough to assume an efficient correction of the noise, at the condition that the LLM is accurately prompted (context, goal of the normalization and list of phenomenons to correct). However, the results of the sequence labeling task do not confirm this hypothesis. Some semantic information may be lost and lead to a negative impact on the sequence labeling task. In the example in Figure 1, we can see a case where the LLM overcorrected and modified a critical word. The word "DEFECTUEUX" (tr. "defective") was replaced by "défaut" (tr. "defect"). In the reports, "defect/défaut" is often found and used

[2] https://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/

[3] https://www.gnu.org/software/wdiff/

with the meaning of "problem" or "inadequacy". As such, it has been manually labeled most of the time with the label "Out of specification". It differs from "defectueux/ defective" which means that a component is not functioning, thus labeled with "Not working". In this example, by changing this particular word, the LLM has modified the meaning of the sentence and even its correctness (in this case "est défaut" is nor grammatically accurate, nor attested in the corpus). The classifier applied on the rectified text thus incorrectly labels "est défaut" as "Out of specification", while the original text get a correct label of "Not Working" for "DEFECTUEUX". To conclude, we can say that the use of transformers models on noisy and technical data seems to be quite robust and able to cope with such a corpus, addressing the main types of noise. However, the noise itself does not seem to bear the difficulty of the sequence labeling task given the score obtained on the normalized dataset close to the score of the original reports dataset.

Acknowledgments

The authors would like to thank the CNES and MeetSYS for funding this work and providing the corpus and expertise required for this study.

Experiments presented in this paper were carried out using the OCCIDATA platform that is administered by IRIT and supported by CNRS and University of Toulouse (https://occidata.irit.fr).

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).

Tyler Bikaun, Melinda Hodkiewicz, and Wei Liu. 2024a. 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 K. Bikaun, Tim French, Michael Stewart, Wei Liu, and Melinda Hodkiewicz. 2024b. MaintIE: A fine-grained annotation schema and benchmark for information extraction from maintenance short texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10939–10951, Torino, Italia. ELRA and ICCL.

Quinten Bolding, Baohao Liao, Brandon Denis, Jun Luo, and Christof Monz. 2023. Ask language model to clean your noisy translation data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3215–3236, Singapore. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and Aurelien Rodriguez. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Pierre Falzon. 1987. Langages opératifs et compréhension opérative. *Le Travail Humain*, n°50:281–286.

Loïc Galand, Michal Kurela, and Horacio Romero Clavijo. 2018. Techniques de TAL pour la recherche des "signaux faibles" et catégorisation des risques dans le REX SDF des lanceurs spatiaux. In *Congrès Lambda Mu 21, " Maîtrise des risques et transformation numérique : opportunités et menaces "*, Reims, France.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2627–2638, New York, NY, USA. Association for Computing Machinery.

Michal Kurela, Mathilde Bacqué, and Remi Laurent. 2020. Classification automatique des faits techniques pour la conformité des lanceurs spatiaux. In *Congrès Lambda Mu 22 " Les risques au cœur des transitions " (e-congrès) - 22e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement, Institut pour la Maîtrise des Risques*, Le Havre (e-congrès), France.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, 41(3):437–479.

Isaac Persing and Vincent Ng. 2009. Semi-supervised cause identification from aviation safety reports. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 843–851, Suntec, Singapore. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 101–108.

Ludovic Tanguy, Nikola Tulechki, Assaf Urieli, Eric Hermann, and Céline Raynal. 2016. Natural language processing for aviation safety reports: From classification to interactive analysis. *Computers in Industry*, 78:80–95.

Alan Thomas, Robert Gaizauskas, and Haiping Lu. 2024. Leveraging LLMs for post-OCR correction of historical newspapers. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 116–121, Torino, Italia. ELRA and ICCL.

Bin Wang, Chengwei Wei, Zhengyuan Liu, Geyu Lin, and Nancy F. Chen. 2024. Resilience of large language models for noisy instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11939–11950, Miami, Florida, USA. Association for Computational Linguistics.

Chuyang Yang and Chenyu Huang. 2023. Natural language processing (nlp) in aviation safety: Systematic review of research and outlook into the future. *Aerospace*, 10(7):600.

James Zhang, Wouter Haverals, Mary Naydan, and Brian W. Kernighan. 2024. Post-ocr correction with openai's gpt models on challenging english prosody texts. In *Proceedings of the ACM Symposium on Document Engineering 2024*, DocEng '24, New York, NY, USA. Association for Computing Machinery.

A Prompts

Included in prompt version	Text
1,2,3,4	You are a trained linguist working with maintenance operators. Your task is to correct sentences written in French by these operators. These texts describe problems occurring during the maintenance of a rocket. You are correcting these texts because they contain a lot of noise. You must write a standardized version of these texts without modifying, reformulating, or changing any words. Do not alter the vocabulary.
2,3,4	You need to clean these text because they will be automatically processed afterward.
3,4	Here is a list of the different phenomenons to correct you may encounter : - missing spaces and punctuation - mispelled words - the whole text in uppercase - missing accents. Even if you encounter an unfamiliar word, keep it as it is. When displaying your answer, write only the corrected version of the sentence without adding line breaks, additional information, explanations, or notes.
4	Here are two examples. The text "CORROSION LEGERE SUR OVM50005CORROSION PLUS IMPORTANTE SUR OVM5006 (VANNE ALIM PISCINE)" becomes "Corrosion légère sur OVM50005. Corrosion plus importante du OVM5006 (vanne ALIM piscine)." The text "POULIE DE RENVOI SUR CAISSSON LBS LH2 GRIPPEE SUR SON AXE." becomes "Poulie de renvoi sur caisson LBS LH2 grippée sur son axe.".
1,2,3,4	Here is the text to rectify: [text inserted]