Natural Language Processing for Risk, Resilience, and Reliability

Jean Meunier-Pion

Natural Language Processing for Risk, Resilience, and Reliability

Jean Meunier-Pion[1]

[1]*CentraleSupélec Université Paris-Saclay, Gif-sur-Yvette, Essonne, 91190, France*
jean.meunier-pion@centralesupelec.fr

ABSTRACT

Natural Language Processing (NLP) has seen a surge in recent years, especially with the introduction of transformer architectures, relying on the now famous self-attention mechanism. Especially, with the rise of Large Language Models (LLM), propelled by the appearance of ChatGPT in 2022, a new hope of extracting relevant information from text has emerged. In the meantime, natural language data have not often been used in risk, resilience, and reliability tasks. However, text data containing reliability-related information, that can be used to monitor health information regarding complex systems, are available in several and diverse shapes. Indeed, text data can either contain theoretical expert knowledge (technical reports, documentation, Failure Modes and Effects Analysis (FMEA)), or in-practice expert knowledge (incident reports, maintenance work orders), or in-practice non-expert knowledge (customer feedback, news articles). Critical infrastructures, such as nuclear powerplants, railway networks, or electrical power grids, are complex systems for which any failure would induce severe consequences affecting many people. Such systems have the advantage of serving many users, thus having many possible text sources from which technical information and past incident data can be mined for anticipating future failures and generating responses to catastrophic scenarios. The goal of this work is to develop methods and apply state-of-the-art NLP techniques to text data relating to critical infrastructures and failures, to (1) mine information from unstructured language data, and (2) structure the extracted information. Preliminary experiments were conducted on customer review data and incident reports, and show promising performance for failure detection from text data with transformers, as well as incident-related information extraction using LLMs.

1. STATEMENT OF THE PROBLEM ADDRESSED

Risk, resilience, and reliability have seen some attempts to use Natural Language Processing (NLP) to make use of text data in systems health monitoring. NLP was applied to maintenance records data so as to filter maintenance records by types (Stenström et al., 2015). Sharp et al. (2017) also developed a framework on maintenance records, to classify such data based on expert tags and by supervised learning. Considering the specificity of technical terms used in maintenance work orders, Brundage et al. (2021) introduced the notion of Technical Language Processing (TLP) and discussed the need for models designed and trained specifically on technical language data. Other works (Li & Wu, 2018; Huang et al., 2021) have proposed a statistical approach to look at co-occurrences of terms and a graph visualization to quickly perceive how failures are characterized in diesel engines, based on Failure Modes and Analysis Effects (FMEA) data. Research in NLP for risk, resilience, and reliability covers multiple applications, with different datasets and tasks. However, it suffers from a lack of common shared open-source datasets and benchmarks, and with the rise of generative artificial intelligence, there is currently room for improving existing frameworks and developing new ones.

The research question addressed in this thesis is the following: how can one extract information from unstructured text data, and then structure the extracted information, to learn failure knowledge from text data?

The initial approach should involve using state-of-the-art NLP techniques, especially Large Language Models (LLMs) and transformers (Vaswani et al., 2017) in general, to extract information from text. The extracted information will then be organized in knowledge databases, according to ontologies, in order to structure the information relating to risk, resilience, and reliability. The goal is to use the large amount of available text data containing health information of complex systems so as to learn and structure knowledge on failures of critical infrastructures.

To that end, various forms of text can be used. Either documents containing theoretical expert knowledge, such as technical reports, technical documentation, or FMEA; or documents with in-practice expert knowledge, such as incident reports, or maintenance work orders; or documents with in-practice non-expert knowledge, such as customer feedback, or news articles. Such data can then be used in two complementary ways: either to directly mine information

from them, or to give context and knowledge while extracting information from other documents.

The expected benefits are the creation of tools in the form of specialized technical search engines and automated text assistants to support informed decision-making for the anticipation of incidents and the generation of response scenarios when encountering failures in critical infrastructures.

2. Expected Contributions to the Field

The main expected contributions to the field include (1) the development of open-source datasets to support NLP tasks applied to risk, resilience, and reliability, (2) the application of state-of-the-art NLP techniques, including LLMs to reliability data and the creation of associated benchmarks, (3) the design of an ontology for reliability engineering and the development of a method to automatically populate knowledge databases whose architecture would rely on this ontology.

3. Research Plan

The research plan currently includes the following parts: (1) detecting failures and assessing reliability from text data, (2) applying LLMs for information extraction, (3) focusing on failure mode extraction with the proposed framework for information extraction assisted by LLMs, (4) designing an ontology for reliability engineering, and (5) automatically populating knowledge databases for system reliability.

As a transversal task, the development of fine-tuned LLMs for risk, resilience, and reliability tasks, e.g., including code generation for reliability engineering, is a common thread.

3.1. Failure Detection and Reliability Assessment from Text Data

The simplest unit of information that can be extracted from text data regarding reliability is whether or not the document at hand states that a failure occurred.

Following previous research (Meunier-Pion et al., 2021), a set of customer review data for failure detection was developed for the task of detecting if customers report a failure in their review of a product. It is composed of 2,415 customer reviews labeled for binary classification. Additionally, labels include a level of granularity that enables the subtask of classifying failures severity as tolerable or intolerable.

Due to the ambiguity of customer reviews, several annotators were required to label the dataset and a human benchmark score was derived from the annotations to know what the best performance of a machine model could be. The human performance was estimated to 91.24% of balanced accuracy, while the best model involving a fine-tuned DeBERTa-v3 transformer (He et al., 2023) reached 88.50% balanced accuracy. This constitutes promising results for detecting failures in customer review data, and in natural language in general, in order to generate lifetime data from text corpora and assess reliability directly from natural language data.

The results from this research part suggest that reliability-related information can be extracted from text data. Building upon this preliminary work, the aim of this thesis is to gather more fine-grained information regarding systems health, such as failure causes, failure modes, degradation, maintenance actions, interdependencies between system components, and so on.

3.2. Application of LLMs to Information Extraction

With the rise of LLMs and their incredible capabilities for understanding natural language, it seems that NLP information extraction tasks can be addressed more effectively. However, one limitation of LLMs is that they are designed for generating text, in the form of long consecutive sentences, instead of returning only a specific word or set of words answering a short query.

In this research, LLMs were applied on nuclear powerplants incident reports data for extracting basic information such as the date of an incident and the place of an incident. Using a small LLM stored on less than 3 GB, an average accuracy of 94.5% could be reached for the extraction of date and place of incidents, over an initial dataset of 50 incident reports. Besides, one should note that if a LLM outputs “The date of the incident was 2023.”, then the output is considered invalid, as the expected queried information is only “2023”, making the task more challenging as conciseness matters.

The goal is to provide a framework for extracting information thanks to LLMs, that combines the ability of LLMs to understand text and generate high quality answers, with a methodology for extracting specific queried information. Here, in this part of the research, the goal is not necessarily to extract technical information, but rather to come up with an effective and performant framework for extracting pre-defined attributes when queried, as illustrated in Figure 1.

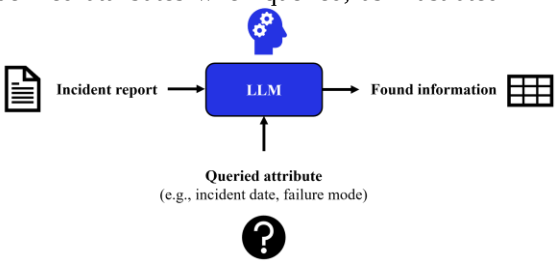

Figure 1. Information extraction using an LLM.

3.3. Failure Mode Extraction

By leveraging the framework developed for Section 3.2., one type of reliability-related information that can be queried

from text is the failure mode of a given system, i.e., how the system failed. As part of this research, methods are developed to extract failure mode information from text.

There are mainly two ways of extracting failure mode information from text: (1) classifying failure mode based on pre-defined failure mode labels, or (2) generating a failure mode label that fits the given description of an incident.

While the first way, involving multi-class classification may be convenient to assess performance and compare models, it requires the definition of class labels, which are not always available, especially when working with new unseen data. On the other hand, the second envisioned approach to extracting failure mode from text involves generating labels, which requires a more sophisticated evaluation framework. This approach is intended in this research, in order to leverage LLMs for information extraction. Additionally, one modern technique that will be investigated in this research is Retrieval-Augmented Generation (RAG), which consists of generating an answer to a query, with the addition of a context from a vector database and similar to the input query.

In the meantime, the first way of extracting failure mode information is currently under study and preliminary results on the National Highway Traffic Safety Administration (NHTSA) complaints dataset show that it is possible to reach 86% balanced accuracy on multi-class classification of failure modes on text data, using only standard NLP techniques, without even the use of transformers.

3.4. Definition of an Ontology for Reliability

The objective of this research being to learn failure knowledge from text data, one important part of this work is to define an ontology for reliability. Previous works in maintenance have already applied ontology frameworks to define ontologies like an Ontology model for Maintenance Strategy Selection and Assessment (OMSSA) (Montero Jiménez et al., 2023).

The purpose of defining an ontology for reliability is to organize concepts relating to failures in order to structure failure knowledge. This should enable and facilitate the automatic instantiation of knowledge databases containing failure information extracted from text data.

3.5. Automatic Population of Knowledge Databases

Ultimately, the purpose of this research is to enable the automatic population of knowledge databases containing failure-related information extracted from text data.

In that respect, a challenge that will be addressed in this research is grouping fields of the same data record. Indeed, multiple data records can have their information in the same document and an additional challenge thus is: how to distinguish between different data records? How can one group fields together to create the correct instance, and not mix fields from different records together?

More specifically, fields that can be extracted from text data include, for example, the date of an incident, the failure mode, and the root cause of the failure. The challenge is to correctly map the date of incident A with the failure mode and root cause of A, and not map it with the failure mode and root cause of B, whenever A and B co-occur in a document.

3.6. Fine-Tuning LLMs for Reliability Engineering

As part of this research, a transversal component will be the development of fine-tuned LLMs specialized on technical data in order to efficiently use technical engineering data and to address tasks relating to system health monitoring.

In that respect, a first attempt of benchmarking LLMs on the fields of risk, resilience, and reliability, is under study and involves the creation of a dataset for code generation containing more than 50 code generation questions. This dataset is inspired by the HumanEval dataset (Chen et al., 2021) and involves the usage of unit tests to guarantee the capability of the model to generate effective code. The goal is to evaluate current state-of-the-art LLMs, such as variations of Mistral or Llama models, on the vertical application of risk, resilience, and reliability, whereas traditional code generation benchmarks (Austin et al., 2021; Du et al., 2023) consist of general programming tasks.

Then, an LLM will be fine-tuned on specific data to compensate for the lack of expert knowledge from general LLMs, and enable the generation of more accurate technical scripts from an artificial intelligence code assistant. This approach will be generalized to fine-tune LLMs not only for code generation, but also for natural language in general, in order to acquire expert knowledge on complex systems.

4. CONCLUSION

The current research aims at developing open-source datasets and benchmarks for NLP for risk, resilience, and reliability, while leveraging state-of-the-art techniques like LLMs. The main focus here is the development of methods for information extraction and structuring knowledge.

Preliminary results show encouraging evidence that state-of-the-art NLP techniques are able to mine failure-related information from text data. Nonetheless, the methods developed in this thesis are intended to be applied to critical infrastructures, thus confidence indicators are necessary to measure the trustworthiness of the developed models.

As a common thread, an objective throughout this research is to create NLP-related materials, including datasets and code, that will be shared to encourage research in this field and ensure access to trustful and reproducible results.

REFERENCES

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., & Sutton, C. (2021). *Program Synthesis with Large Language Models*. Preprint. Google Research.

Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., & Lukens, S. (2021). Technical language processing: Unlocking maintenance knowledge. *Manufacturing Letters*, vol. 27, pp. 42-46. doi: 10.1016/j.mfglet.2020.11.001

Chen, M., et al. (2021). *Evaluating Large Language Models Trained on Code*. Preprint. OpenAI, San Francisco, California, USA.

Du, X., Liu, M., Wang, K., Wang, H., Liu, J., Chen, Y., Feng, J., Sha, C., Peng, X., & Lou Y. (2021). *ClassEval A Manually-Crafted Benchmark for Evaluating LLMs on Class-level Code Generation*. Preprint. Fudan University, Shanghai, China.

He, P., Gao, J., & Chen, W. (2023). *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing*. Preprint. Microsoft Azure AI.

Huang, Q., Wu, G., & Li, Z. T. (2021). Design for Reliability Through Text Mining and Optimal Product Verification and Validation Planning. *IEEE Transactions on Reliability*, vol. 70, pp. 231-247. doi: 10.1109/TR.2019.2938151

Li, Z., & Wu, J. (2018), A Text Mining based Reliability Analysis Method in Design Failure Mode and Effect Analysis. *2018 IEEE International Conference on Prognostics and Health Management (ICPHM)*. June. doi: 10.1109/ICPHM.2018.8448909

Meunier-Pion, J., Zeng, Z., & Liu, J. (2021). Big Data Analytics for Reputational Reliability Assessment Using Customer Review Data. *Proceedings of the 31st European Safety and Reliability Conference (ESREL 2021)*. September 19-23, Angers, France. pp.2336-2343

Montero Jiménez, J. J., Vingerhoeds, R., Grabot, B., & Schwartz, S. (2023). An ontology model for maintenance strategy selection and assessment. *Journal of Intelligent Manufacturing*, vol. 34, pp. 1369-1387. doi: 10.1007/s10845-021-01855-3

Sharp, M., Sexton, T., & Brundage, M. P. (2017). Towards Semi-autonomous Information: Extraction for Unstructured Maintenance Data in Root Cause Analysis. *IFIP International Conference on Advances in Production Management Systems*. March 9, London, England. pp.425-432

Stenström, C., Aljumaili, M., & Parida, A. (2015). Natural Language Processing of Maintenance Records Data. *International Journal of Condition Monitoring and Diagnostic Engineering Management*, vol. 18, pp. 33-37.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017), Attention is All you Need. *Advances in Neural Information Processing Systems*.