CIRP Annals - Manufacturing Technology 72 (2023) 33–36

Contents lists available at ScienceDirect

CIRP Annals - Manufacturing Technology

journal homepage: https://www.editorialmanager.com/CIRP/default.aspx

Automation of knowledge extraction for degradation analysis

Sri Addepalli[a], Tillman Weyde[b], Bernadin Namoano[a], Oluseyi Ayodeji Oyedeji[a], Tiancheng Wang[b], John Ahmet Erkoyuncu (2)[a,*], Rajkumar Roy (1)[b]

[a] School of Aerospace, Transport systems and Manufacturing ad materials (SATM), Cranfield University, Cranfield, United Kingdom
[b] School of Science & Technology, City, University of London, London, United Kingdom

ARTICLE INFO

Article history:
Available online 15 June 2023

Keyword:
Knowledge management
Decision making
Knowledge graph

Degradation analysis relies heavily on capturing degradation data manually and its interpretation using knowledge to deduce an assessment of the health of a component. Health monitoring requires automation of knowledge extraction to improve the analysis, quality and effectiveness over manual degradation analysis. This paper proposes a novel approach to achieve automation by combining natural language processing methods, ontology and a knowledge graph to represent the extracted degradation causality and a rule based decision-making system to enable a continuous learning process. The effectiveness of this approach is demonstrated by using an aero-engine component as a use-case.

1. Introduction

Complex engineered high value assets (HVA) such as aero-engine components tend to require a high level of maintenance due to the complex functionality and the fact that they are made up of multiple complex systems that serve a singular purpose, in this case to power the aircraft to fly. The modern jet engine is a fine example that demonstrates an all-round complexity, be it in its design, manufacturing, service or end-of-life. The maintenance of such a HVA is a significant undertaking especially when the activities are centred around reliability and safe operation of the engine, which encompasses a large portfolio of parts and systems that need extensive maintenance support during its operational life. Rising competition and stringent regulations have become major decision control points pushing manufacturers and operators into achieving high levels of reliability. The aerospace industry is currently undergoing digital transformation and adopting digital business models to improve efficiency. With the industry's reliance on documentation where data is presented in the form of a text-based reports, the last decade has seen a huge move towards paperless record and book-keeping, with a vast amount of historical information still in hard records and/or digital records that do not contain any meta-data identifying what the information in those documents is [1]. To review and extract knowledge manually at this scale becomes extremely laborious. In order to effectively use and manage the information, suitable digital tools, applying natural language processing (NLP) are being developed to extract knowledge from the copious amounts of documented information so that fully informed decisions (including probability of failure) can be made.

Degradation analysis has been of pivotal interest for decades, where the degradation of a part or system can be identified, measured and directly related with the potential failure or the remaining life of that part. Traditional and advanced inspection systems belonging to the family of Non-destructive Testing (NDT) have been adopted to characterise and manage damage occurring in parts that could potential reduce functionality or lead to complete failure of the asset [2]. For a number of years, the aerospace maintenance sector has been collecting information about the relationship between degradations, their effects and root causes, primarily to understand the remaining useful life of the asset. This information in the form of text based reports, still does not exist in a comprehensive digital form and is embedded as tacit knowledge within the organisations that use this information. Over the last few decades, organisations have been collecting copious amounts of data that have been stored in digital formats without proper classification and exists in fragments as reports across their database. This paper thus proposes a novel end-to-end digital platform that improves and effectively automates knowledge extraction and representation of textual databases with bespoke functionality tailored to the end-user, in this case the maintainer. We focus here on the textual data and make no attempt to use or extract numeric information or use numeric knowledge. We use NLP and machine learning to not just automatically extract information from the database(s) hosted within an organisation but also to recognise causal links.

2. NLP in the context of degradation analysis

Text-mining using NLP has attracted great attention in industry as it helps process thousands of documents and extract information that might otherwise be labour-intensive when compiled manually [3]. Specifically, extracting causal relationships can help populate Knowledge Graphs (KG) with relevant knowledge on multiple levels, providing the user with navigable information representations. Informed decisions can significantly improve reliability and reduce resource wastage in line with methodologies such as 'lean manufacturing' [4].

* Corresponding author.
E-mail address: j.a.erkoyuncu@cranfield.ac.uk (J.A. Erkoyuncu).

https://doi.org/10.1016/j.cirp.2023.03.013

A variety of traditional NLP techniques such as pattern-based extraction, clustering, latent semantic analysis, co-occurrence and contrastive analysis exist. However, the challenge with these still is that whilst they can produce satisfactory results, they require high levels of expert intervention combined with high computational cost leading to challenges when applying them to large real-world scenarios. Over the last decade, deep learning has become the dominant technique for knowledge extraction with the pre-trained language models becoming the standard.

Recurrent neural networks, based on Long Short Term Memory (LSTM) [5], such as ELMO [6], have largely been replaced by attention based Transformer architectures [7], such as BERT [8] and OpenAI's GPT models [9]. In this, context words (or sub-word units) are represented as vectors that can be learned separately using Skip-gram [10], Continuous Bag Of Words (CBOW) [11], Contextualized Word Vectors (CoVe) [12], Global Vector (GloVe) [13], and Fast Text [14], or as part of a model. We apply deep learning models in this work to extract relevant terms and causal relations from text.

3. The knowledge extraction

For effective automation, various requirements were captured from practioners and included both functional requirements, such as the ability to provide the information in the form of a triple (degradation root cause, mechanism, effect), and metrics that show the context and/or occurrence including the exact location of the particular occurrence in the database. As part of the requirements, a degradations taxonomy was initially created manually and was used as a benchmark to perform a semi-supervised machine learning activity to enable knowledge extraction from the database.

This paper proposes an automated knowledge extraction framework demonstrated and validated with a corpus of published papers on degradations occurring in Nozzle Guide Vanes (NGV) (a critical aero-engine component) as a use-case. In this paper, the framework has been designed to enable the use of an end-to-end knowledge-base system that indexes all the information and present it to an end-user as seen in Fig 1.

To elucidate the framework (Fig 1), the elements of the framework are explained below

Query – This is where an end-user inputs the context-specific degradation term they are interested in.

Search – The query initiates the web-scraping activity that searches the database containing the search term. In this particular case, databases, such as SCOPUS, ACM, IEEE and ARXIV, have been used. These databases' include sources to a wide variety of academic publications as indicated in Fig 1. The search was aimed at only the metadata in this case and included items such as title, keywords and abstract. For the purpose of the demonstration only papers with metadata were considered.

Filtering – A pre-processing step was included to filter data that was not relevant to the context. In this instance, the context was set to NGV and only those papers that included NGV and degradations and or the terminologies established using the taxonomy were considered. This enabled effective filtering of disparate and non-relevant data thereby excluding them from our corpus.

Download – This was the last step in the web-scraping process where full academic papers were downloaded directly to a data repository using the filtered data.

Data indexing – In this step, Named Entity Recognition (NER) was applied to identify and index of the collected data. In order to perform the indexing activity, the following concepts and elements were used to effectively improve the indexing operation.

- *Sentence array* –Sentences are represented as arrays of word or sub-word tokens. In order to compare and contrast, spans (sub-arrays) are used to identify multiple phrases of relevance in the text.
- *Causal connective* – In order to extract the relationship automatically, a causal connective semantic is introduced, i.e., a word or phrase that indicates causality. Typical examples are 'because of', 'is caused by', 'due to', 'results in', 'lead to' and so on. Identifying correct spans of the connective and the related cause and effect (see below), are needed for correct degradation identification.
- *Cause and Effect* – These are spans in a sentence that the causal the connective in-between refers to. With this approach we can identify variable numbers of degradation terms within the same sentence. However, due to the large dataset involved in this work, only those sentences that contain two or more degradation terms and a causal connective are evaluated.
- *Named Entity Recognition (NER)* – The taxonomy in the context of the use-case, in this example NGVs, has been used to perform the NER that identifies the terms of interest within a text using a token-based classification system with each token assigned to a specific label for better identification.

Causality Extraction – The first step in performing the causality extraction is the creation of the taxonomy based on expert knowledge and public information, validated with industry and academic experts (also subject matter expert (SME)). Based on multiple discussions, 3 industry-academia workshops conducted over a period of 6 months, a causality model based on triples was first established and used as a baseline for this activity (Fig 2).

In this step the relationship extraction is undertaken by expanding the NLP activity which is as follows

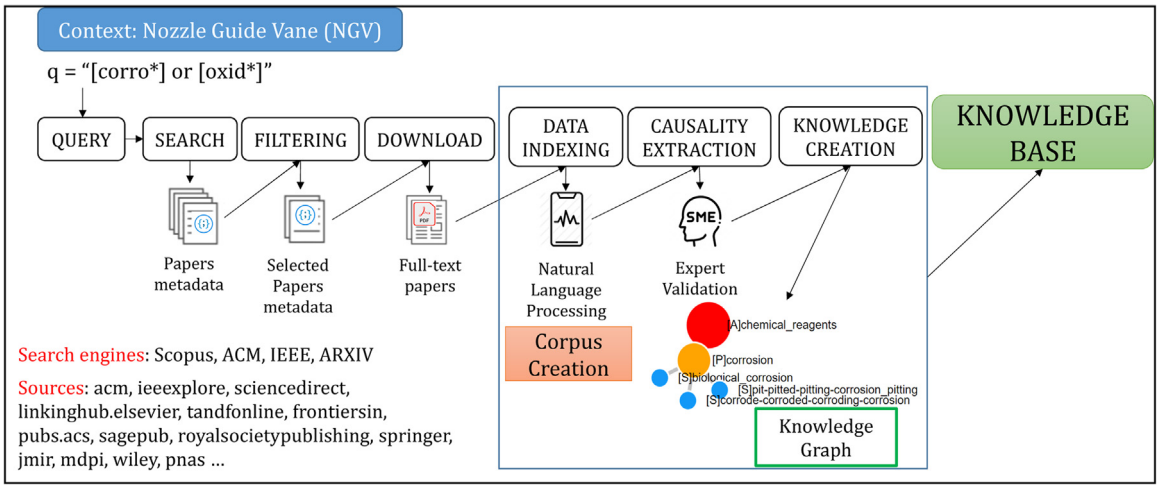

Fig. 1. The automated knowledge extraction framework.

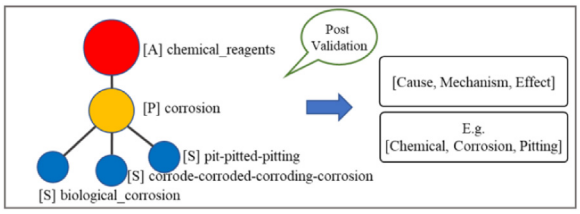

Fig. 2. Causality mapping activity (left) showing the triples with an example (right).

- *Relationship Extraction (RE)* – An established RE model [15] for binary relationship extraction has been used. For instance, if the sentence is constructed as

$$S = w_1, w_2, .., e_1, \ldots, w_j, ..e_2, ..w_n \quad (1)$$

Where $w_1..w_n$ are the words, e_1 & e_2 are the entities, then the mapping function can be reproduced as;

$$F_R(T(S)) = +1, If e_1 and e_2 are related \quad (2)$$

and

$$F_R(T(S)) = -1, Otherwise \quad (3)$$

This establishes if the entities in the sentence have a relationship or not. For the purpose of this research, the contiguous causal relation recognitions or CCRR and a BEERT model have been used to extract the causality from the data.

- *Model Architecture* – In the current setup, the entire document is input, read with the document's metadata, title, abstract and the main body of the paper. Based on the Taxonomy, the text containing the degradation information is first extracted and broken down into sentences and tokens. We match the text to the Taxonomy terms allowing for spelling mistakes, word splitting and merging multiple verb forms in English. For the purpose of this paper, the Bidirectional Encoder Representations from Transformers (BERT) [8] has been adopted to perform both terminology recognition and relationship extraction. BERT is a transformer-based model architecture that is trained to predict masked words. BERT includes all of the contextual information including the terms of interest and their relationship in the pretraining stages. A fully-connected feed-forward layer was applied on top of the BERT model for classification.
- *Evaluation metrics* – In order to evaluate the model, Precision, Recall and F1 score are used. Based on True Positives (TP) and False Positives (FP) and False Negatives (FN), the precision, recall and F1 scores are calculated as follows:

$$precision = |TP|/(|TP| + |FP|) \quad (4)$$

$$recall = |TP|/(|TP| + |FN|) \quad (5)$$

$$F1 = (2 \times precision \times recall)/(precision + recall) \quad (6)$$

The F1 metric balances False Positives against False Negatives. Further, to achieve higher degrees of matching the Jaccard Index has been used prior to performing the F1 approximations and the following is the expression

$$J(A, B) = |A \cap B|/|A \cup B| \quad (7)$$

Knowledge Representation – In order to present the information, an app has been developed to present the extracted information on a user-friendly interface. In order to make sure that the triples' relationship is maintained at high levels throughout the activity, the ontology was modelled initially only at the highest class level (cause, mechanism and effect). In order to allow efficient transfer of knowledge and to keep the model terminologies correct, three equivalent classes were defined (attribute, property and state), which are aligned to the original triples.

Fig 3 is a snapshot of the degradation ontology devoloped using protégé. This ontology supports the Graphic User Interface (GUI) toolkit that is used to present the outputs to the end-user.

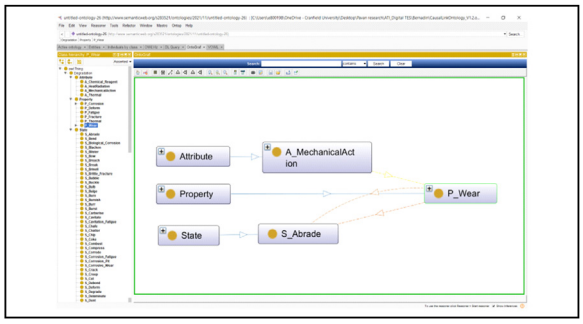

Fig. 3. A snapshot of the high-level degradation ontology.

- *Knowledge graph (KG)* – Using the existing concepts of knowledge representations [4] in the form of a graph, the following process has been adopted to represent the extracted information and is as seen in Fig 4. The process for this specific KG creation started with the creation of the corpus, filtering and cleansing the data, NER tagging, word tokenization, application of BERT, extracting triples and then populating the KG and presenting the results to the user.

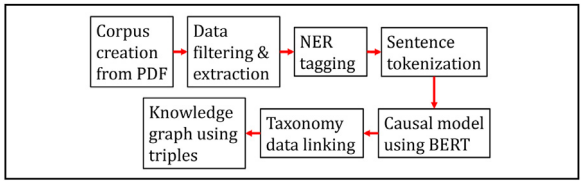

Fig. 4. The overall knowledge extraction process.

Knowledge base – The output from the data mining activity gets added to the knowledge base. This becomes the final repository of all the information captured during the entire process. In this case, this step is included in the framework to show that the outputs of this knowledge extraction activity feeds back into the knowledge base improving the digital data management activities of the organisation.

3.1. Results from the case study

To demonstrate the framework, publicly available academic journal and conference papers were used, that relate to the keywords "NGV" and "degradation". The web-scraping activity, from the filtering stage of the framework produced an initial 6000+ research papers in PDF format specific to NGVs and degradation. Through the stated methodology of filtering presented above (Section 2 – Filtering), only research papers containing the context and the required information were included and this resulted in 1237 papers in the final corpus.

Using the BERT model and the taxonomy data (Table 1) the RE activity to populate the output in triples was conducted and the following was achieved (Table 2).

It should be noted that, at the point of defining the taxonomy certain fundamental terminology changes were made. Characteristics

Table 1
Excerpt from the populated taxonomy (including verb forms and synonyms).

Attribute	Property	State 1	State 2	State 3
Mechanical Action Chemical Reagent	Wear Corrosion	abrade corrode	abraded corroded	abrading corroding

Table 2
Output from the BERT model.

Sentence: "Spillage of concentrated HCL caused the product to deform"		
Attribute	*Property*	*State*
Mechanical Action, chemical reagent	Corrosion – Wear	Deform

such as temperature, pressure, moisture, particulate etc. have been reported sparingly as a root cause and in most cases as a combinational attribute. In order to streamline the outputs, higher level classifications for the root causes have been produced (e.g., mechanical action, chemical reaction, heat/radiation etc.). The outputs in Fig 5 were also represented into a KG (Fig 6).

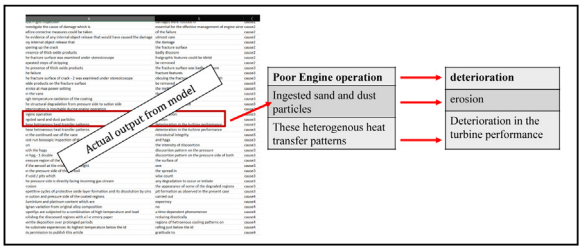

Fig. 5. Sample output from the RE activity populating results in triples.

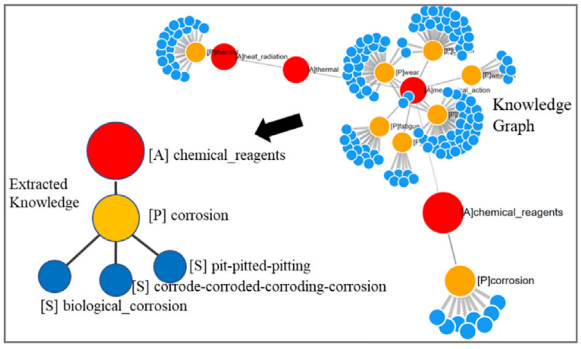

Fig. 6. The outputs presented in the form of a KG represent content from Fig 5.

3.2. Model evaluation

As indicated in the Evaluation Metrics section above (Section 3 pg3), the model evaluation was conducted on the token classification and our combined token and span classification method. It can be seen clearly that the combined method yielded higher F1 scores (see Section 3 – Evaluation Metrics, pg3) for the connective, cause, effect and the whole sentence, as indicated in Table 3. The result shows a higher Recall in the case of the connective significantly improving the extraction of the terms associated with degradation (90% against 64%). This allowed us to filter the data and extract only those instances that contained the connective, the cause, the effect and the combination of all three scenarios.

Table 3
Model evaluation results indicating the performance of the algorithms.

	Token Classification			Our Method		
	Precision	Recall	F1	Precision	Recall	F1
Connective	65%	64%	64%	68%	90%	77%
Cause	61%	63%	62%	62%	72%	67%
Effect	68%	71%	70%	70%	76%	73%
Whole Sentence	42%	75%	54%	59%	70%	64%

It should be noted that, these higher scores represent improved overall system efficiency in line with the results obtained. The outputs in the form of the KG (in excess of 20 graphs created from the corpus) were also checked manually, validated against the taxonomy and found to be much superior than the outputs from the token based method.

4. Discussions and conclusions

The novelty of this paper is in the proposed approach for automation in knowledge extraction with the notion of improving decision making especially in the context of degradation assessment and maintenance planning. Whilst academic research tends to focus on methods to create ontologies and KGs, it becomes very important to show the steps and know-how of building automated systems that aim to achieve efficiencies in addition to disruptive digital transformations in the HVA environment.

This paper thus shows how an automated knowledge extraction system can work on a real-world use case by offering a direct approach to building digital systems that support the human-in-the-loop by providing the end user with both current and historical information in a meaningful format enabling a fast and robust decision-making process. The approach of combining NLP, machine learning and knowledge representation presented in this paper shows superior quality of the degradation knowledge being extracted, which is supported by the higher F1 scores presented in Table 3. Further, the outputs not just improve the overall assessment allowing continuous learning growing the degradation knowledge but also becomes the content for digital tools such as digital twins that are currently being built to support HVAs operation and management. This work introduces causal modelling in an engineering information system, ontology-based vocabulary expansion for NER applied to degradation processes in aero-engineering to mine maintenance data. For future work, the architecture is proposed to be tested in further use-cases evaluating the applicability across sectors. There is also a need for evaluating different types and scales of databases to conduct text mining.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is supported by the Innovate UK project titled: Digitally Optimised Through-life Engineering Services [DOTES] with the grant no. 103082–263289. Data supporting this study are not publicly available due to commercial reasons. Please contact the corresponding author.

References

[1] Adams C., 2020, Going Paperless in the Hangar, https://www.avm-mag.com/going-paperless-in-the-hangar/ (accessed Jan. 13, 2023).
[2] Fortune Business Insights, 2022, Airport Systems /Non-Destructive Testing (NDT) Market, (accessed: Jan. 13, 2023)
[3] Dogra V, et al. (2022) A Complete Process of Text Classification System Using State-of-the-Art NLP Models. *Computational Intelligence and Neuroscience*. https://doi.org/10.1155/2022/1883698.
[4] Liu A, et al. (2022) Knowledge Graph with Machine Learning for Product Design. *CIRP Annals* 71(1):117–120.
[5] Melamud O, et al. (2016) Context2vec: Learning Generic Context Embedding with Bidirectional LSTM. *20th SIGNLL Conference*, 51–61. https://doi.org/10.18653/v1/K16-1006.
[6] Peters M.E. et al., 2018, "Deep contextualized word representations,".
[7] Vaswani A. et al., 2017, "Attention is all you need," Advances in Neural Information Processing Systems (NIPS2017)
[8] Devlin J. et al., 2018, "BERT: pre-training of Deep Bidirectional Transformers for Language Understanding,".
[9] Radford A. et al., 2018, "Improving Language Understanding by Generative Pre-Training,".
[10] Mikolov T. et al., 2013, "Distributed Representations of Words and Phrases and their Compositionality,".
[11] Mikolov T. et al., 2013, "Exploiting Similarities among Languages for Machine Translation,".
[12] McCann B, et al. (2017) Learned in Translation: contextualized Word Vectors. *Advances in Neural Information Processing Systems*, v. 30.
[13] Pennington J, et al. (2014) GloVe: Global Vectors for Word Representation. *EMNLP2014*, 1532–1543. https://doi.org/10.3115/v1/D14-1162.
[14] Bojanowski P. et al., 2016, "Enriching Word Vectors with Subword Information,".
[15] Bach N, Badaskar S (2007) A Review of Relation Extraction. *Literature Review for Language and Statistics II* 2:1–15.

2023-07-13

Automation of knowledge extraction for degradation analysis

Addepalli, Sri

Elsevier

Addepalli S, Weyde T, Namoano B, et al., (2023) Automation of knowledge extraction for degradation analysis. CIRP Annals - Manufacturing Technology, Volume 72, Issue 1, July 2023, pp. 33-36

https://doi.org/10.1016/j.cirp.2023.03.013