Towards Learning Causal Representations of Technical Word Embeddings for Smart Troubleshooting

Alexandre Trilla[1,3], Nenad Mijatovic[2], and Xavier Vilasis-Cardona[3]

[1] *Alstom, Santa Perpètua de la Mogoda, Barcelona, 08130, Spain*
alexandre.trilla@alstomgroup.com

[2] *Alstom, Saint Ouen, Paris, 93482, France*
nenad.mijatovic@alstomgroup.com

[3] *DS4DS, La Salle, Universitat Ramon Llull, Barcelona, 08022, Spain*
xavier.vilasis@salle.url.edu

Abstract

This work explores how the causality inference paradigm may be applied to troubleshoot the root causes of failures through language processing and Deep Learning. To do so, the causality hierarchy has been taken for reference: associative, interventional, and retrospective levels of causality have thus been researched within textual data in the form of a failure analysis ontology and a set of written records on Return On Experience. A novel approach to extracting linguistic knowledge has been devised through the joint embedding of two contextualized Bag-Of-Words models, which defines both a probabilistic framework and a distributed representation of the underlying causal semantics. This method has been applied to the maintenance of rolling stock bogies, and the results indicate that the inference of causality has been partially attained with the currently available technical documentation (consensus over 70%). However, there is still some disagreement between root causes and problems that leads to confusion and uncertainty. In consequence, the proposed approach may be used as a strategy to detect lexical imprecision, make writing recommendations in the form of standard reporting guidelines, and ultimately help produce clearer diagnosis materials to increase the safety of the railway service.

1. Introduction

Natural Language Processing (NLP) provides an effective approach for improving the collection and analysis of text-based maintenance data, and eventually enable accurate decision-making (Brundage, M. P., Weiss, B. A., and Pellegrino, J., 2020). For example, in the railway maintenance business, axle bearings are some of the most critical rolling stock components subject to strong safety constraints. In consequence, many conservative overhaul actions are scheduled preventively in the maintenance plan, which contains a lot of technical documentation about these mechanical assets. The completion of these actions, in turn, generates useful practical feedback on the shop floor following the inspection of the parts, which seeks degradation signals and compiles them in written maintenance sheets. Additionally, unexpected failures like grease leaks, hot axleboxes, or abnormal vibration records, get reported in an issue tracking system to be then fixed correctively. Considering all these environments together entails dealing with a large amount of text data that is oftentimes manually intractable, and NLP brings the automation potential to extract useful insights to advise the maintenance team, e.g., by identifying the most probable underlying root cause to a given problem. This approach is meant to increase the chances of success to fix the issue, minimize the risk of a recurrent failure, and thus maximize the availability of the fleet.

Interactive natural language interfaces help maintainers achieve a higher success rate and a lower task completion time, which lead to greatly improved user satisfaction (Su, Y., Awadallah, A. H., Wang, M., and White, R. H., 2018). However, many solutions require customization through the collaboration between data scientists and domain specialists, and each technical field poses its own challenges. In this sense, Technical Language Processing (TLP) presents a holistic, domain-driven approach, to use NLP in a technical engineering setting (Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., and Lukens, S., 2021). In TLP, maintenance documents like work orders are relatively small in size and contain misspellings, domain-specific jargon, abbreviations, and non-standard sentence structure. Therefore, to tackle this particular context-dependent technical scenario, the field of causality is

https://doi.org/10.36001/IJPHM.2022.v13i2.3127

regarded as a direct description of what occurs when machines degrade, and the root-cause analysis becomes the means to obtain a reliable troubleshooting explanation for an abnormal failure. In fact, linguistic representation, such as the one found in TLP, is essentially a causal phenomenon (Stampe, D. W., 2008).

Causality is traditionally stratified into a three-layer hierarchy (Pearl, J., 2019): association (i.e., plain correlation or direction-free relationships), intervention (i.e., reasoning about the effects of actions), and counterfactuals (i.e., retrospective reasoning). In turn, Causal Inference (CI) aims to draw such detailed interpretations beyond mere associations from observational data using statistical tools to infer relational probabilities. CI distinguishes two broad classes of causal queries: forward causal questions or the estimation of "effects of causes", and reverse causal inference or the search for "causes of effects" (Gelman, A., and Imbens, G., 2013). CI can also be conceptualized as a multitask learning problem with a set of shared layers among the factual and counterfactual outcomes (Alaa, A. M., Weisz, M., and van der Schaar, M., 2017). Similarly, decision-making is about predicting counterfactuals (Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M., 2017), and CI can potentially lead to more informed decisions (Zheng, M., Marsh, J. K., Nickerson, J. V., and Kleinberg, S., 2020). The difficulty here is that all these probabilistic quantities are not directly available in observational/factual data, so the CI problem needs to be converted into a domain adaptation problem to figure out the mechanisms that explain why observations occurred (Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A., 2020).

Understanding causality is considered as one of the current challenges for Machine Learning (ML) automation because ML models are ultimately driven by correlations in the data, and in general the causality implications of interest cannot be derived from them (Ahmed, O., Träuble, F., Goyal, A., Neitz, A., Bengio, J., Schölkopf, B., Wüthrich, M., and Bauer, S., 2020). Therefore, counterfactual explanations are gaining prominence as a way to explain the decisions of a ML model (Barocas, S., Selbst, A. D., and Raghavan, M., 2019). The causality hierarchy, and the formal restrictions it entails, explains why ML systems can attain CI as long as they model the data beyond mere observed associations. Therefore, learning causal relations can be transformed into a supervised prediction problem once the data labels indicate the causal directionality, whether explicitly or implicitly (Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H., 2020; Shalit, U., Johansson, F. D., and Sontag, D., 2016). In this line of work, research in ML and language understanding have recently found a great deal of success using large neural networks, especially through Deep Learning (DL) (Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N., and Fox, E. A., 2020; LeCun, Y. and Bengio, Y., and Hinton, G. E., 2015). These overparameterized and regularized models constitute one of the most important ideas in the recent history of statistics, along with CI (Gelman, A., and Vehtari, A., 2020), and a straightforward way to learn causal effects and counterfactual outcomes with DL is to learn representations for features, i.e., to let the DL system automatically discover the most effective way to represent the data directly instead of hard-coding traditional language features. To this end, DL-based word embeddings may provide an interesting approach to represent linguistic causality (Li Y., and Yang T., 2018; Hancock, J. T., and Khoshgoftaar, T. M., 2020).

Specifically, Word Embeddings (WE) are dense, fixed-length word vectors, built using word co-occurrence statistics as per the distributional hypothesis (Almeida, F., and Xexéo, G., 2019). WE learn representations of high-level abstract concepts of the kind humans manipulate with language, away from the perceptual space, and they exhibit some geometric relational properties (Bengio, Y., 2017), which can ultimately be used to conduct lexical comparisons (Tan, L., Zhang, H., Clarke, C. L. A., and Smucker, M. D., 2015). Thus, this data representation can be regarded as an approach to cognition and artificial intelligence (Maguire, P., Mulhall, O., Maguire, R., and Taylor, J., 2015). Moreover, WE are computationally efficient (Levy, O., and Goldberg, Y., 2014), and therefore they need less data to successfully train statistical models (Goth, G., 2016), as is the case in TLP. Regarding semantics, WE also expose word senses (Yaghoobzadeh, Y., Kann, K., Hazen, T. J., Agirre, E., and Schütze, H., 2019), but they may experience the meaning conflation deficiency that arises from representing a word with all of its possible meanings as a single vector (Camacho-Collados, J., and Pilehvar, M. T., 2018). Nevertheless, WE constructed using arbitrarily contextualized language have further improved representational performance, possibly helping in the semantic disambiguation of machine decay (Levy & Goldberg, 2014; Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L., 2018). In this line, WE also lead the way to process language in Prognostics and Health Management (PHM) because they display a high flexibility that is only attained by avoiding task-specific engineered features (Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., and Ducoffe, M., 2020).

The troubleshooting objective pursued in this article is interesting for the PHM community to enhance the maintenance business (Leao, B. P., Fitzgibbon, K. T., Puttini, L. C., and de Melo, G. P. B., 2008). Realizing a comprehensive monitoring of system data, a timely detection of system abnormalities, and troubleshooting are all worthy goals, and the recent exponential growth of PHM patents is a point of support for these advantages (Liu, Z., Jia, Z., Vong, C.-M., Han, W., Yan, C., and Pecht, M., 2018). Current troubleshooting tools rely on fault tree analysis, extensive electronic manuals or expert system methods to assist the maintainer in identifying faulty system components (Naveed, A., Li, J., Saha, B., Saxena, A., and Vachtsevanos, G., 2012). The approach presented in this paper combines these complementary methods through the exploitation of technical text data from different environments, which is aligned with the scope of PHM (Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., and Lukens, S., 2021).

This work applies the CI paradigm to PHM using DL through a contextualized WE to better troubleshoot the root causes of failures and help improve their diagnostics. To do so, it exploits two different linguistic environments where causality is expected to be observed. On the one hand, an ontological reference framework based on a Failure Mode, Mechanism, and Effect Analysis (FMMEA), which provides a scholarly structure of causality driven by degradation. On the other hand, an actual record on Return On Experience (ROX), the data of which have been explicitly written for the purpose of explaining the root causes of the reported failures. In both environments, several experts inherently identify which properties of the observations describe spurious correlations unrelated to the causal explanation of interest, and which properties represent the phenomenon of interest, i.e., the stable invariant correlations (Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D., 2019). In this controlled analysis dealing with experimental data, invariant correlation implies causation. Therefore, DL and WE should be adequate tools to extract the textual regularities that represent causality (Mitrovic, J., McWilliams, B., Walker, J., Buesing, L., and Blundell, C., 2021), and thus they may be used to rate the level of agreement between CI theory and practice for troubleshooting. Specifically, a probabilistic Causality-Contextualized WE (CCWE) is trained with the ROX data, and the FMMEA-based failure ontology data is then used to evaluate the alignment between the two environments, which is expected to be reasonably high. This hypothesis is validated experimentally using the technical documentation related to rolling stock bogies. Figure 1 shows a diagram of the proposed analysis workflow for clarity.

The article is organized as follows: Section 2 describes the data, i.e., the bogie FMMEA and ROX records, the way the ontology has been created, and the strategy to build a CCWE. Section 3 conducts a graphical analysis of the whole failure network to discover structurally interesting points, a probabilistic analysis of the ROX-based CCWE to assess the causal relationships in practice, and the integration of the two perspectives, including a distributed representation of causality. Section 4 discusses the limitations of the proposed approach through the comparison with an alternative spectral embedding and the modeling of textual sequences. Finally, Section 5 concludes the manuscript showing how the concept of causality in bogie failures has been partially attained with the current technical documentation, and how it may be improved with the approach presented in this work.

2. MATERIALS AND METHODS

In this section, a FMMEA for bogies is used to build a failure ontology of their degradation, and a text database of ROX data is used to build a practical CCWE.

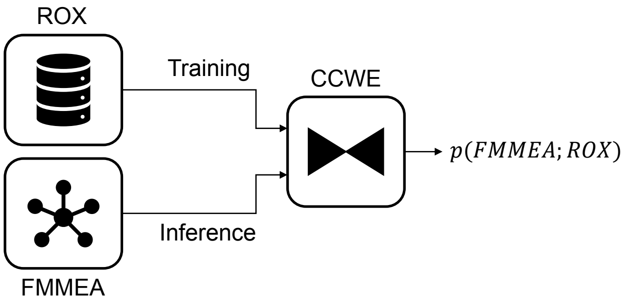

Figure 1. Diagram of the probabilistic analysis workflow performed in this work, which evaluates the level of agreement between two causality-rich environments: the Failure Mode, Mechanism, and Effect Analysis (FMMEA) on the theoretical side, and the Return On Experience (ROX) on the practical side. A Causality-Contextualized Word Embedding (CCWE) is developed to model and evaluate the relevant causal linguistic regularities.

2.1. Failure Ontology

The FMMEA is an efficient tool to analyze system and component failures, and identify their main causes or mechanisms of failure (Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., and Zerhouni, N., 2017). Knowledge of the failure mechanisms that are likely to produce the degradation that can lead to eventual failures in the monitored assets is important to succeed in the implementation of a PHM solution (Mathew, S., Das, D., Rossenberger, R., and Pecht, M., 2008). Therefore, the FMMEA is one of the tools used for the effective assessment of risk, and so it is a vital part of an organization's strategic management. However, it is costly to produce and hardly reusable due to its text-based description in natural language (Ebrahimipour, V., Rezaie, K., and Shokravi, S., 2010). To overcome this situation, an ontology-based solution is advised to extract and reuse FMMEA knowledge from the available text documents (Rehman, Z., and Kifor, C. V., 2016).

An ontology is a network of standard concepts and terms in a given domain that shows their properties and the relations between them to represent knowledge (Ebrahimipour, V., Rezaie, K., and Shokravi, S., 2010). There is a growing interest in the potential value of ontologies to codify structures of meaning for maintenance (Sexton, T., Hodkiewicz, M., Brundage, M. P., and Smoker, T., 2018). To this end, TLP is the way to go to automatically extract valuable insights regarding the many facets of reliability, maintenance, and planning (Navinchandran, M., Sharp, M. E., Brundage, M. P., and Sexton, T. B., 2019). The ontology augments human decision-making by relying on diversified information (Polenghi, A., Roda, I., Macchi, M., and Pozzetti, A., 2022), especially when real-life maintenance data is used in its design. Conforming to the vocabulary that is widely used by maintenance professionals and practitioners is a major catalyst for widespread acceptance and uptake (Karray, M. H., Ameri, F., Hodkiewicz, M., and Louge, T., 2019). Additionally, to tackle CI with the ontology, its topology needs to represent a Structural Causal Model (SCM) framework because its organization is essential for performing causality learning tasks such as counterfactual reasoning (Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y., 2021). Explicitly, a SCM consists of a set of explanatory variables, outcome variables, and unobserved variables, connected by a set of functions that determine their relational values (Pearl, J., 2009).

For the analysis of bogie failures framed in this work, a FMMEA approach is recommended to reduce blindness, subjectivity, and over-reliance on the personal experience (Li, Y.-H., Wang, Y.-D., and Zhao, W.-Z., 2009). And for the successful application of CI, assumptions about the mechanisms underlying the observed data also need to be specified (Sharma, A., and Kiciman, E., 2020). To this end, the approach proposed by Atamuradov and colleagues is taken for reference in this work, and thus its contents are not questioned here (Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., and Zerhouni, N., 2017). Their failure analysis defines three fields that are described as follows, along with the related causal structure:

Failure Mechanism Fundamental manner in which a component can fail → Unobserved variable that is the Root Cause of an observed Problem, e.g., fatigue or wear

Failure Mode Manner by which a failure is physically observed, although in certain contexts, the Failure Effect (i.e, the impact of the Mechanism) can also be found in this field → Outcome variable that represents a Problem that is experienced, e.g., surface defects, rotation difficulty, or reduction of suspension effect

Component Explanatory variable that describes the context of a Problem, e.g., wheel or gearbox

Components are related to Failure Modes, which in turn are then related to Failure Mechanisms. If these relationships are likened to an ISO 13379 standard causal tree with faults, symptoms, and descriptors (ISO, 2003), the resulting failure ontology is shown in Figure 2, where the directed edges indicate the (assumed) direction of causation (Imbens, G. W., 2020).

2.2. Return On Experience

ROX is a holistic approach to understand and increase the value of investments across customer, employee, and leadership experience (PwC, 2019). It is strictly related to the First Time Right management principle, which aims to minimize the number of product issues that get past design release and cause rework, leading to dissatisfied customers (Leuenberger, H., Puchkov, M., and Schneider, B., 2013). Specifically, ROX is a data-driven quality strategy that focuses on identifying and eliminating the root cause of the problems and ensure that the improvement is sustained (Smetkowska, M., and Mrugalska, B., 2018). To this end, tagging and curating already existing textual data can be a first step toward structuring content (Sexton, T. B., and Brundage, M. P., 2019), but this work goes beyond this step and processes data that have been specifically written for the purpose of describing the causal sources of the reported problems. Therefore, unlike regular observational data, ROX records are hardly marred by selection biases, confounding factors, and other such weak points, and thus they may be treated as experimental or interventional data.

The ROX database of use in this work contains around 500 records written by many experts following a feasible collaborative approach (Hastings, E. M., Sexton, T., Brundage, M. P., and Hodkiewicz, M., 2019). However, different technicians rarely describe the same Problem in an identical manner or register (Conrad, S., 2019). This leads to description inconsistencies within the database and makes it difficult to categorize issues or learn from similar causal relationships (Sharp, M. E., Sexton, T. B., and Brundage, M. P., 2017). Therefore, a statistics-based TLP approach is needed to put the focus on factual data and strip grammatical artifacts, e.g., by filtering out stop words, lemmatizing, etc. This provides a systematic methodology to create computable knowledge (Sexton, T., Hodkiewicz, M., Brundage, M. P., and Smoker, T., 2018).

By definition, plain text data are intrinsically unstructured. However, in the ROX database each record conducts a specific troubleshooting analysis in isolation, and the causal connections are organized into the following fields:

Problem Subject title, description of the reported Failure Mode, and details of its technical impact.

Root Cause Description of the Failure Mechanism of the issue following an investigation, and the main reason of non-detection.

Business context Strategic unit: trains, rail services, rail control, and infrastructure.

System context Technical scope: air supply, passengers, roof, door, and bogie.

Issue context Domain category: mechanical, documentation, electrical, and assembly.

Table 1 shows some examples of bogie ROX database entries to illustrate the nature of these data (note that the majority of the instances are mechanical issues).

To further understand the characteristics of these technical text data, which justifies the TLP-based approach, Figure 3 shows the power-law distribution of its ranked word frequencies compared to what is expected in natural language (Zanette, D. H., and Montemurro, M. A., 2005). Note that the technical language curve has a positive offset with respect to natural language. This increased word frequency spectrum may indicate that this technical language shows a reduced vocabulary and

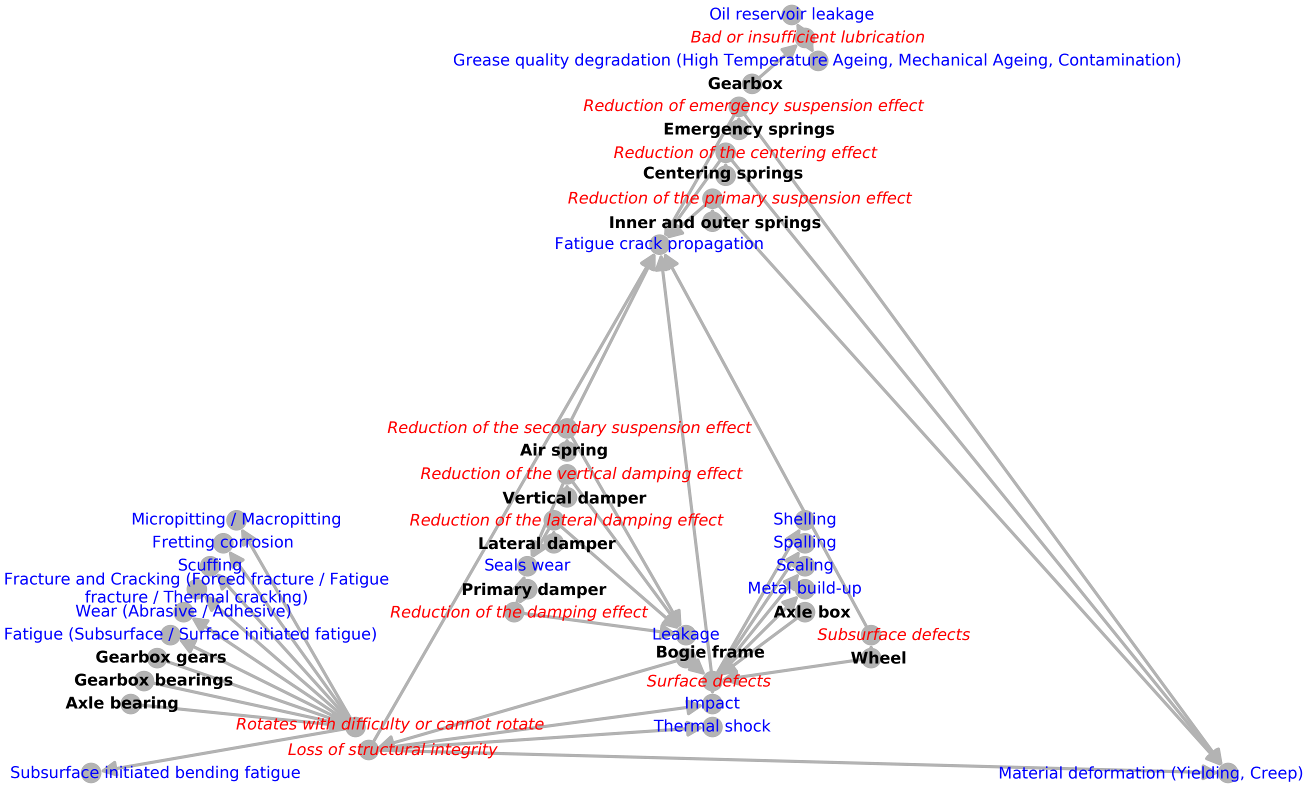
Oil reservoir leakage
Bad or insufficient lubrication
Grease quality degradation (High Temperature Ageing, Mechanical Ageing, Contamination)
Gearbox
Reduction of emergency suspension effect
Emergency springs
Reduction of the centering effect
Centering springs
Reduction of the primary suspension effect
Inner and outer springs
Fatigue crack propagation
Reduction of the secondary suspension effect
Air spring
Reduction of the vertical damping effect
Vertical damper
Micropitting / Macropitting
Reduction of the lateral damping effect
Shelling
Fretting corrosion
Lateral damper
Spalling
Scuffing
Seals wear
Scaling
Fracture and Cracking (Forced fracture / Fatigue fracture / Thermal cracking)
Primary damper
Metal build-up
Wear (Abrasive / Adhesive)
Reduction of the damping effect
Axle box
Fatigue (Subsurface / Surface initiated fatigue)
Leakage
Subsurface defects
Bogie frame
Gearbox gears
Wheel
Surface defects
Gearbox bearings
Impact
Axle bearing
Rotates with difficulty or cannot rotate
Thermal shock
Loss of structural integrity
Subsurface initiated bending fatigue
Material deformation (Yielding, Creep)

Figure 2. FMMEA-based bogie failure ontology linking Component (black boldface) to Failure Mode (red italics) and then to Failure Mechanism (blue).

therefore the same words may need to be used more often. In a similar descriptive vein, Figure 4 shows the distribution of technical ROX text lengths as word counts per record along with some comparative hints regarding natural language. Note that the statistical ROX length mode is around 8 words, which is far from the optimum contemporary readability indication of 17 words (DuBay, W. H., 2004). Such short texts have some unique characteristics that make them difficult to handle. For instance, they do not always observe the syntax of written language, they contain limited context, and they give rise to ambiguity as more than one meaning may be conveyed, leading to vagueness and confusion (Wang, Z., and Wang, H., 2016). Moreover, the ROX length distribution shows a tail of longer texts that get increasingly difficult to read, and also over 35 words the quality of a language model decreases rapidly (Bahdanau, D., Cho, K., and Bengio, Y., 2015).

2.3. Causality-Contextualized Word Embedding

The original conception of a WE related a single word to its local context given a shallow window of proximity (Mikolov, T., Chen, K., Corrado, G., and Dean, J., 2013). However, this principle does not hold for CI because the context of the related texts is different. In this work, the goal is to learn the causal relationships between Problems (i.e., Failure Modes) and their Root Causes (i.e., Failure Mechanisms) through their respective textual expressions. To do so, a binary-valued Bag-Of-Words (BOW) model is considered to account for the presence of multiple words concurrently (Le, Q., and Mikolov, T., 2014). Note that the syntax is not retained as this model focuses on the overall semantics through the lexicon. In turn, the input and output vocabularies are also dependent on their causal roles, and regarding that an effective method depends on the size of the vocabulary (Chen, W., Grangier, D., and Auli, M., 2015), both Root Cause and Problem lexicons are considered in the present WE model.

The proposed implementation of the CCWE for troubleshooting is based on an encoder-decoder DL architecture using the causal concept of refinements (Mitrovic, J., McWilliams, B., Walker, J., Buesing, L., and Blundell, C., 2021), see Figure 5. Root Causes are probabilistically modeled given their Problems and some Context, which is a situational hint to enhance language models (Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., and Jiang, M., 2020), and may be stripped from the model once trained. The CCWE is exploited with a contrastive esti-

Table 1. ROX database examples of bogie system failures reported by maintenance services.

Issue context	Problem	Root Cause
mechanical	vertical damper failure, sealing defect	as per supplier investigation report the failure mode is the primer glue departed from the metal parts. considers it was because the metal parts were not cleaned well while in the pre treatment process the primer glue can't adhere well to the metal parts so it will cause debonding issue during operation.
mechanical	anti roll bar assembly knocking noise, excessive noise	a light stick slip phenomena is the root causes of the noise. it is decided to change the knuckle as per updated design from supplier hyed for one complete train set. currently in claim situation with supplier for them parts are compliant to specification.
mechanical	oil leakage from gear box unit, loss of tightness	as per supplier rca it is confirmed that the gear lubricating oil from the drainage hole leakage caused by the labyrinth ring tw of roundness error. oil leakage causes in the process of the part in ngc in sheet2 the process of operation not suitable for the mode of transportation easy to cause roundness error of deformation when parts fall off or pressure deformation.
assembly	conical spring bonding issue, loss of regulation	debonding beetwen rubber material and steel frame incorrect handling of adhesived parts by operators before putting them into the mould. the cleanliness of localalized area is jeopardized and it disturb the bonding process between rubber and interface. it was not possible to detect during the validation tests the parts tested did not presented failure. the issue happens when submitted to load sometimes with few milage or more than 150.000 km for example.

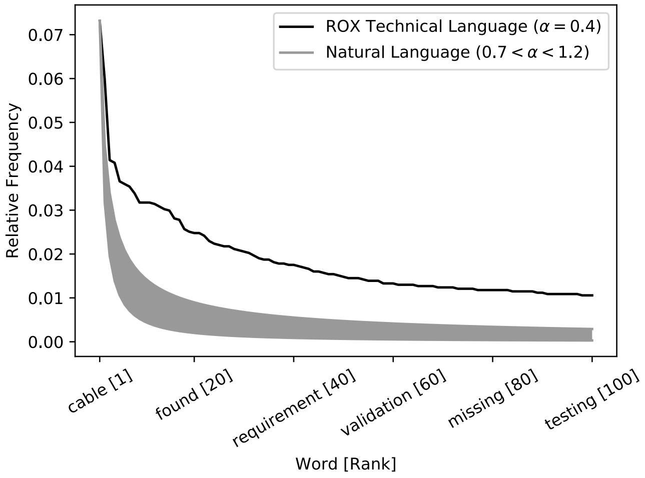

Figure 3. Ranked word relative frequency distribution of technical ROX text data versus natural language. The exponent of the power laws is shown in brackets.

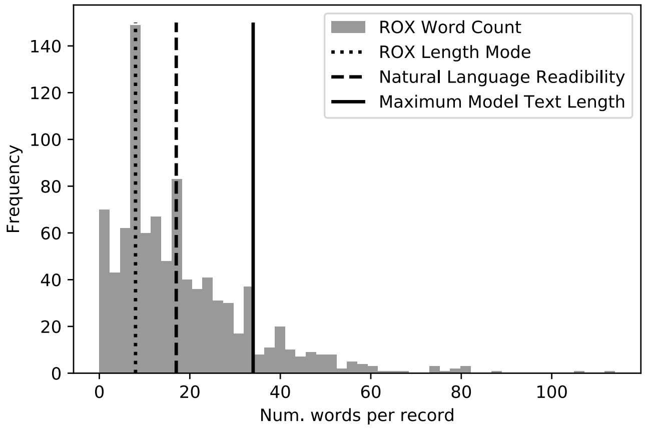

Figure 4. Word count frequency distribution of technical ROX text data records along with natural language readability indications.

mation framework, which discriminates between the observed data and some artificially generated noise (Gutmann, M., and Hyvärinen, A., 2010; Mnih, A., and Teh, Y. W., 2012; Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J., 2013). This approach is attained through jointly learning a series of nonlinear logistic regressions using an output logistic activation function and a cross-entropy cost criterion for training. Bias terms are also considered because of the multiple-word instances with different lengths (there is no basis to assume that the embedding will be centered around the origin). Also, being a DL solution the model is expected to be overparameterized, so the use of Dropout layers is recommended to manage words that belong to regions of poor overlap in the feature space (Alaa, A. M., Weisz, M., and van der Schaar, M., 2017). Specifically, the input layer is followed by a Dropout layer to deal with long texts because these are more likely to have words deactivated, therefore equaling their potential impact to that of shorter instances. And the embedding layer, which is smaller than the BOW-based layers, is also followed by another Dropout layer to adjust its representational expressiveness and manage ambiguity more effectively (Yaghoobzadeh, Y., Kann, K., Hazen, T. J., Agirre, E., and Schütze, H., 2019).

The proposed CCWE model gives the following probability directly:

$$p(Root\ Cause|Problem, Context)$$

However, an explicit formulation through the embedding bottleneck layer is advantageous to study the geometric properties of its distributed representation, see Eq. (1).

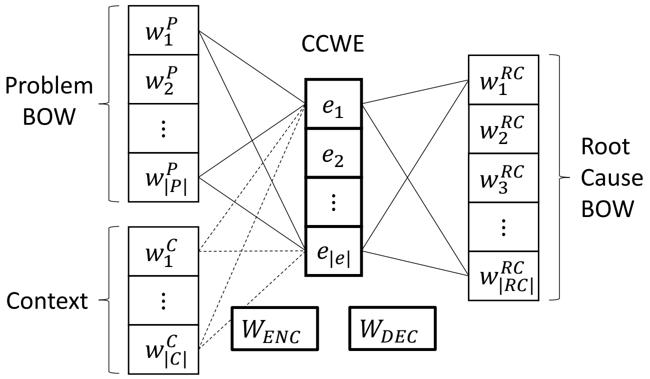

Figure 5. Encoder-decoder DL architecture of the CCWE in inference mode. Dropout layers are used in training mode only, and are thus not shown here for clarity.

$$\begin{aligned} CCWE &= W_{ENC} \cdot (Problem, Context) \qquad (forward) \\ &\sim W_{DEC}^{+} \cdot logit\,([\,Root\,Cause\,]\,) \quad (backward) \end{aligned} \tag{1}$$

Note that the backward equation requires the inversion of the non-square decoder matrix W_{DEC}, which is not possible. In this case, a least-squares approximation is used through its pseudoinverse W_{DEC}^{+}. Also note that the $logit$ function cannot be applied to a binary-valued BOW vector because it leads to an asymptotic overflow. In this case, the values of the $[\,Root\,Cause\,]$ vector are clipped to 0.2 (false) and 0.8 (true). These bounds are driven by the extrema of the second derivative of the logistic function and prevent its saturation.

Finally, the distributed representation of causality is to be exploited through the Principal Components (PC) of the CCWE and the cosine distance between Root Cause and Problem BOW vectors (Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J., 2013). The angle they form in the PC space is a common textual similarity metric utilized in semantic classification and search (Tan, S., Zhou, Z., Xu, Z., and Li, P., 2019). And taking into account that the cosine similarity becomes less predictive as the dimensionality increases (Yaghoobzadeh, Y., Kann, K., Hazen, T. J., Agirre, E., and Schütze, H., 2019), the PC representation is typically reduced to two dimensions following the customary practice in NLP research.

3. RESULTS

Causal prediction is not a typical downstream NLP task apt for evaluation. Therefore, the experiments conducted in this section have been compared with human judgments on word relations, i.e., an intrinsic evaluation (Bakarov, A., 2018). Explanations have been provided through graphs, feature importance (e.g., word probabilities), visualizations (e.g., spectral analysis), and concrete examples (Mothilal, R. K., Sharma, A., and Tan, C., 2020).

3.1. Causal Graphs

Graphs are a powerful representation formalism that can be applied to a variety of aspects related to language processing (Mihalcea, R., and Radev, D., 2011). With a proper choice of nodes and edge drawing criteria and weighing, graphs can be extremely useful for revealing regularities and patterns in the data (Nastase, V., Mihalcea, R., and Radev, D., 2015). Additionally, causal graphs reduce the adverse impact of latent variables or noise (Bahadori, M. T., and Heckerman, D. E., 2021). This section studies the failure ontology as a causal graph to detect confounders (i.e., common root causes) as forks, and colliders (i.e., common problems) as inverted forks. To get an overview of these characteristics, centrality measures have been used to pinpoint the most important nodes of the resulting graphs.

On the one hand, the degree centrality $C_D(v)$ states that the important nodes v are the ones that have many connections (Mihalcea, R., and Radev, D., 2011), see Eq. (2), where V is the total number of nodes in the graph, and d is the distance between two nodes, i.e., the minimum number of vertices that separate them. The application of this criterion is shown in Table 2 as a ranking of nodes, and Figure 6 shows a graph that preserves the ontological relationships driven by this ordered arrangement. According to the degree centrality indicator, the confounders are the nodes related to the Failure Modes of the suspension components (i.e., springs, damper...), and the colliders are its Failure Mechanisms (i.e., fatigue crack, material deformation, leakage, and the wear of seals).

$$C_D(v) = \frac{1}{V} \sum_{\forall v\prime \neq v} x, \text{ where } x = \begin{cases} 1 & \text{if } d(v\prime, v) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

On the other hand, the closeness centrality $C_C(v)$ states that the important nodes v are the ones that are near other nodes $v\prime$ (Mihalcea, R., and Radev, D., 2011). This proximity indicator is calculated as the inverse of the sum of the path lengths from a given node to all the other nodes, see Eq. (3). The application of this criterion is shown in Table 3 as a ranking, and Figure 7 shows the corresponding graph that preserves the ontological relationships. According to the closeness centrality indicator, the confounders are the nodes related to the Failure Modes of the bearings: surface defects and rotation difficulty. In general, note that the nodes with the greatest centrality measures are not densely connected among themselves (some even show few connections), thus there are many peripheral items to be taken into consideration.

$$C_C(v) = \frac{V-1}{\sum_{\forall v\prime \neq v} d(v\prime, v)} \tag{3}$$

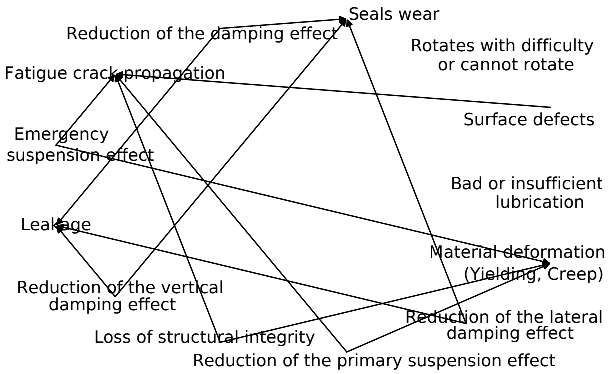

Figure 6. Failure ontology subgraph driven by the nodes with the greatest degree centrality.

Table 2. Ranking of failure ontology nodes according to their degree centrality score.

Failure Ontology Node	Degree Centrality
Rotates with difficulty or cannot rotate	0.2273
Surface defects	0.2045
Fatigue crack propagation	0.1591
Loss of structural integrity	0.1136
Leakage	0.0909
Material deformation (Yielding, Creep)	0.0909
Seals wear	0.0682

Table 3. Ranking of failure ontology nodes according to their closeness centrality score.

Failure Ontology Node	Closeness Centrality
Fatigue crack propagation	0.2121
Leakage	0.1212
Material deformation (Yielding, Creep)	0.1212
Seals wear	0.0909
Impact	0.0710
Surface defects	0.0682
Rotates with difficulty or cannot rotate	0.0682

3.2. Causal Lexical Probabilities

This section conducts a preliminary study of the sensitivity of the CCWE built with the bogie ROX data. The dimensionality of the BOW for the Problem is $|P| = 1591$, for the Root Cause it is $|RC| = 2210$, and for the embedding it is $|e| = 300$. This configuration yields a model with more than 1M trainable parameters. This WE has been trained using cross-validation with a train/test data split of 80%/20%, and the resulting binary accuracy is 0.9894. This learning result indicates that the memorized word relationships of the CCWE are likely to provide reliable causal associations for ROX. To illustrate the troubleshooting capacity of the CCWE, Table 4 shows some generic word examples.

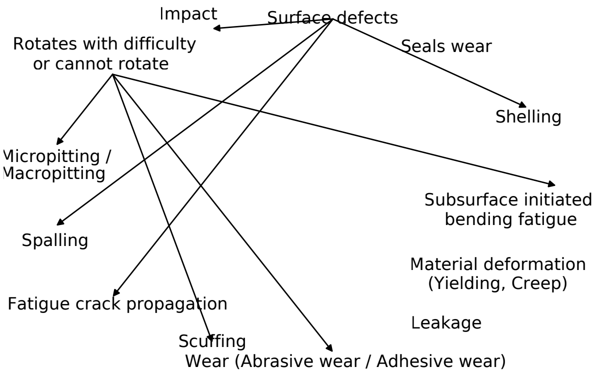

Figure 7. Failure ontology subgraph driven by the nodes with the greatest closeness centrality.

Table 4. Generic troubleshooting word examples obtained with the CCWE.

Problem	Possible Root Cause (Probability)
oil leak	attached (0.8849), measured (0.6137), hole (0.5733), pressure (0.2372)
bearing	tightening (0.0659), vibration (0.0639), shock (0.0495), assembly (0.0394)
gear box	design (0.9062), tolerance (0.9061), oil (0.8703), pressure (0.8237)

In general, the Root Cause outcomes of the CCWE with high probability are reasonable words that belong to the same semantic field of the given Problems. Note that the probabilities for the “bearing” component are an order of magnitude lower than those for “oil leak” and ”gear box”. This result may be due to the specificity of causal words like “tightening”, compared to common words like “attached” or “design”. However, there are also some noise words that typically appear in the BOW of the Root Cause, such as “please”, “report”, “reference”, “part”, etc. This is attributed to the way the experts provide standard ROX feedback. Also, the arrays of output probabilities are mostly comprised of low values, and this is mainly explained by the large space of BOW dimensionality, which leads ROX instances to be sparse.

The geometrical characteristics of the obtained linguistic distributed representation are shown in Figure 8. Note that to obtain this rendering, both the forward encoder and backward decoder equations of the CCWE are needed. This distribution shows that the Root Causes are concentrated in the center, whereas the Problems are spread across the PC space. Thus, the cosine similarity metric is needed to align them within the α angle, yielding a circular sector of causal likelihood. A detailed example of the alignment between a generic Problem like “noise” and its potential Root Causes is shown in Figure 9. The results illustrate the incertitude of the derived causal rep-

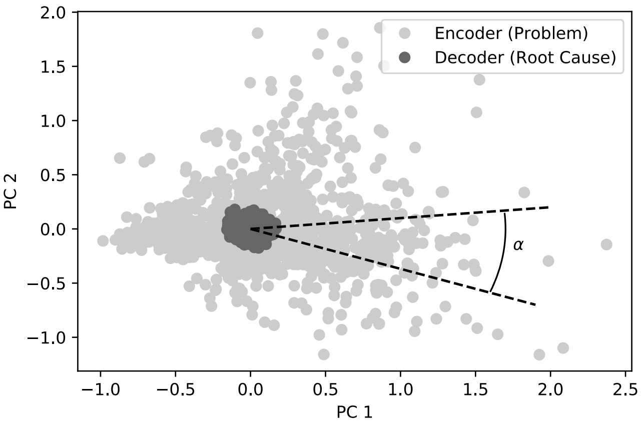

Figure 8. PC of the CCWE activations and the application of the cosine distance similarity measure showing a circular sector α of causal likelihood.

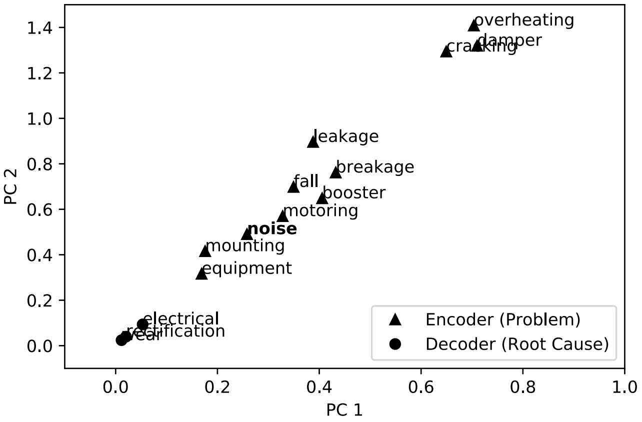

Figure 9. Detailed example of a cosine distance lower than five degrees ($\alpha < 5^o$) between the generic "noise" Problem, its nearest Root Causes, and other close/similar Problems.

resentation as the nearest Failure Mechanisms are ”electrical/rectification" and "wear". In addition, many reasonably related "noise" Problems (sharing the same Root Causes) are also shown, e.g., "motoring", "breakage", "leakage", "cracking", etc.

3.3. Troubleshooting Integration

This section determines if the relationships in the FMMEA-based failure ontology correspond to high ROX-based causal probabilities. To do so, the evaluation of whole Failure Mode texts (as Problems P) is conducted by taking the average probability $\bar{p}_{ROX}(RC|P)$ of the Root Cause RC words appearing in the reported Failure Mechanisms, see Eq. (4), where N represents the words in the text being evaluated. Table 5 shows the top-ranking failures that have been obtained. These results indicate that the leading issues are related to springs and wheels, which the latter is in accord to previous knowledge (Trilla, A., Bob-Manuel, J., Lamoureux, B., and Vilasis-Cardona, X., 2021). Also note that they are mostly linked to the main confounders, i.e., the common root causes, of the failure ontology.

$$\bar{p}_{ROX}(RC_i|P_i) = \frac{1}{N}\sum_{w\in N} p_{ROX}(RC_i^w|P_i) \qquad (4)$$
$$i \in \text{FMMEA Failure Ontology}$$

In addition to this direct FMMEA/ROX relationship, it is also necessary to determine if the cross-failure probabilities are low and thus assert that the proposed approach shows a discriminative property. This alignment study has been determined using the Cross-Probability Difference (XPD) variable, defined by Eq. (5) as the difference between the direct causal probability and the anti-causal probabilities. Note that positive probability differences represent a good alignment between Failure Mode and Mechanism i, whereas negative differences mean that other Failure Mechanisms j are more relevant (according to ROX) than the one stated in the FMMEA-based failure ontology.

$$XPD(i) = \bar{p}_{ROX}(RC_i|P_i) - \bar{p}_{ROX}(RC_j|P_i)$$
$$\forall j \neq i \qquad (5)$$
$$i, j \in \text{FMMEA Failure Ontology}$$

Regarding the distribution of the XPD variable, see Figure 10, the majority of the FMMEA statements are aligned (71.32% of strictly positive values). The clearest textual expressions are driven by the centering springs component. Nevertheless, there are many cases where the difference is too small to extract strong conclusions, as is shown by the high peak around 0. Maybe this is due to averages including missing terms, e.g., specific Failure Mechanism words like "spalling", "scaling", "scuffing", and "pitting" do not appear in ROX. In addition, there are some outlier instances showing a large misalignment, i.e., $XPD < -0.06$. Some examples are listed as follows:

- Bogie frame, Surface defects → Material deformation (Yielding, Creep)
- Bogie frame, Surface defects → Shelling
- Wheel, Surface defects → Material deformation (Yielding, Creep)
- Vertical damper, Reduction of the vertical damping effect → Metal build-up

All these results may be taken for different signs of poor writing, and thus may also be an indication to rephrase those statements and improve the meaning they convey.

To conclude the integration analysis, Table 6 shows an indirect evaluation of the application of the ROX-based causality to the FMMEA-based failure ontology through the cosine distance as the PC vector angle similarity. Bearings and suspension components populate this ranking, which is quite similar to

Table 5. Ranking of ROX-based average probabilities driven by FMMEA failure ontology.

Component	**Failure Mode**	**Failure Mechanism**	$\bar{\mathbf{p}}_{\mathbf{ROX}}$
Centering springs	Reduction of the centering effect	Material deformation (Yielding, Creep)	0.0308
Wheel	Surface defects	Shelling	0.0208
Emergency springs	Reduction of emergency suspension effect	Material deformation (Yielding, Creep)	0.0184
Inner and outer springs	Reduction of the primary suspension effect	Material deformation (Yielding, Creep)	0.0147
Bogie frame	Loss of structural integrity	Material deformation (Yielding, Creep)	0.0071
Bogie frame	Loss of structural integrity	Fatigue crack propagation	0.0021
Bogie frame	Loss of structural integrity	Impact	0.0019

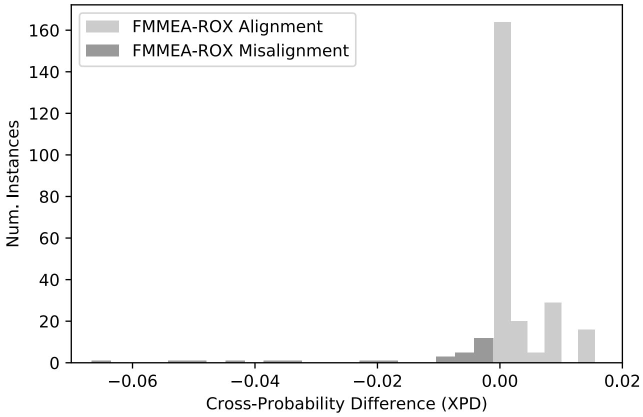

Figure 10. Cross-Probability Difference (XPD) distribution visualized through the histogram.

the one driven by the causal probabilities (a slight reordering is observed, though). In fact, angles and probabilities score a Pearson correlation coefficient of -0.65, so the previous probability-driven conclusions are likely to be largely extrapolated in this causal distributed representation. Therefore, the FMMEA entries that display wide ROX angles may indicate that a rephrasing would be beneficial to increase the comprehension of their text (Ansari, F., 2020). Anyhow, all these results show that the FMMEA ontology relations can be reasonably weighted either via ROX causal probability or distance scores, and thus obtain a SCM to validate the CI approach using a DL-based contextualized WE.

4. DISCUSSION

Up to this point, after having completed the workflow procedure, the discrepancy between FMMEA and ROX has been solely attributed to lexical imprecision between the same causality principles expressed in a particular environment, context, or perspective. However, there may be other sources of epistemic uncertainty that could help explain this divergence. This section addresses some particularities about the proposed CCWE model.

For example, by the Independent Causal Mechanisms principle, the causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other (Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y., 2021). In the troubleshooting probabilistic case tackled in this work, this would imply that the conditional distribution of each Root Cause variable (i.e., the Failure Mechanism) given its Problem (i.e., its Failure Mode) did not inform or influence the other causes. The presented CCWE does not respect this principle because of its multilayer neural topology trained using the standard backpropagation procedure: the encoder layer is influenced by all of the output cause variables, and this, in turn, affects all the predictions through the forward propagation. However, this could also be seen as an advantage from a multitask learning perspective (Crawshaw, M., 2020).

Additionally, performance gains of word embeddings are due to certain system design choices such as dynamically sized context windows and hyperparameter optimizations, rather than the embedding algorithms themselves (Levy, O., Goldberg, Y., and Dagan, I., 2015). This argument leaves the door open to considering *chance* as the ultimate explanatory factor for the results obtained. At the same time, it motivates further research study on DL-based WE.

4.1. Spectral Embedding

Probabilistic models like the CCWE can be viewed as directed graphical models (Salakhutdinov, R., and Hinton, G., 2009). As such, their learned knowledge may be interpreted using a graph spectral embedding or clustering technique. A suitable approach to extract this representation is through the factorization of the Laplacian matrix $L = D - A$, which is a measure of the local derivative of the graph (Mihalcea, R., and Radev, D., 2011). D represents the degree matrix (i.e., the amount of node incoming or outgoing links), and A represents the adjacency matrix (i.e., the causal word relations). After extracting the eigencomponents of L, similar nodes must have embeddings that are close to one another (Cai, H., Zheng, V. H., and Chang, K. C.-C., 2018), and thus the Euclidean distance could be adequate for the similarity comparisons. This section explores this proximity property in the present causal degradation environment.

Figure 11 shows a representation of the two largest Laplacian eigenvectors, which that aim to capture the maximum information (in the form of variance dispersion) of the em-

Table 6. Ranking of ROX-based cosine distance (angle similarity) driven by FMMEA failure ontology.

Component	Failure Mode	Failure Mechanism	α
Gearbox bearings	Rotates with difficulty or cannot rotate	Wear (Abrasive wear / Adhesive wear)	0.5995
Vertical damper	Reduction of the vertical damping effect	Seals wear	5.9321
Axle bearing	Rotates with difficulty or cannot rotate	Wear (Abrasive wear / Adhesive wear)	14.6010
Primary damper	Reduction of the damping effect	Seals wear	15.1803
Centering springs	Reduction of the centering effect	Material deformation (Yielding, Creep)	17.9046
Inner and outer springs	Reduction of the primary suspension effect	Material deformation (Yielding, Creep)	20.4794
Wheel	Surface defects	Metal build-up	23.2299

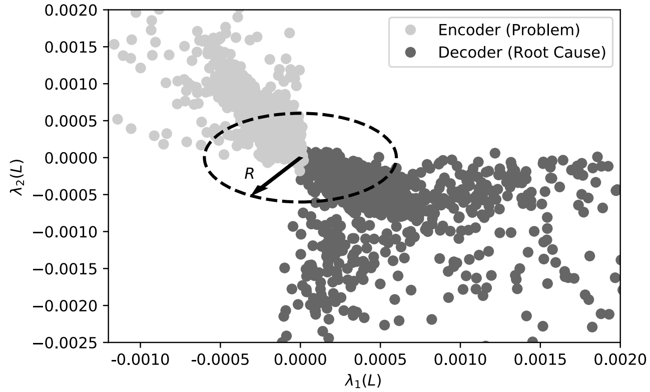

Figure 11. Largest Laplacian eigenvectors λ of the CCWE directed graph and the application of the Euclidean distance similarity measure showing a circle of causal likelihood R.

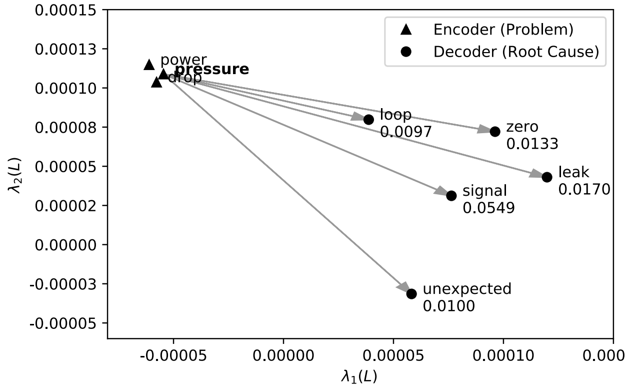

Figure 12. Detailed example of spectral embedding over the generic “pressure” Problem. In this troubleshooting scenario, arrows point toward the potential Root Causes, and the related probabilities are also shown under the words.

bedded causal data. Given the directed bipartite structure of the troubleshooting scenario tackled in this work, where the same word can be used to describe both the Root Cause and the Problem, two degree matrices have been used: one with the Problem word nodes (output degrees only), and the other with the Root Cause word nodes (input degrees only). Finally, their representations have been overlapped, showing that the cause/effect separation is preserved in this low-dimensional illustration. However, only the central region where the two causal roles meet seems to be amenable to any further inference assessment.

Figure 12 shows a more detailed example over the generic “pressure” Problem. All the Failure Mechanism terms that appear seem reasonable given this Failure Mode, e.g., “loop”, “zero”, “leak”, etc. However, in this case, the associated probabilities seem to be unrelated to the distance scores. Moreover, trying to replicate the “noise” Problem used before results in incomprehensible results due to the vast amount of terms that rapidly appear as the radius R is increased. Maybe the factorization of the Laplacian matrix, which is strictly defined for an undirected graph, built over a directed graph is flawed and needs further attention.

4.2. Language Modeling

In previous sections it has been shown that the lexicon per se is sufficient to produce reasonable causal probabilities. However, the principle of semantic composition states that the meaning of a phrase can be derived from the meaning of the words that it contains as well as the syntax that binds them (Iyyer, M., Boyd-Graber, J., Claudino, L., Socher, R., and Daumé III, H., 2014). Likewise, a WE captures syntactic and semantic regularities (Mikolov, T., Yih, W.-t., and Zweig, G., 2013). Consequently, a WE could be able to compose meaningful phrases and thus build a language model.

Language models learn linguistic knowledge, store relational knowledge present in the training data, and may be able to answer structured queries (Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S., 2019). To do so, neural encoder-decoder models pioneered by machine translation were proposed to achieve the goal of mapping input text to output text (Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y., 2014). An encoder network first reads and represents a source sentence into a fixed-length vector, and a decoder network then outputs a target sentence from this encoded vector. This encoder/decoder architecture can also be extended to deal with corpora and vocabulary sizes, and complex, long term structures of language (Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y., 2016). Eventually, encoder and decoder are jointly trained to maximize the conditional probability of a correct relationship, which is conceptually equivalent to what is pursued in the WE but this time considering the sequentiality of words as an additional embedded context (Liu, Q., Kusner, M. J., and Blunsom, P., 2020). This heteroassociative property is explored in this section to relate Root Causes to Problems for long texts.

The specific implementation adopted in this work is based on the Sequence-to-Sequence (S2S) approach. S2S applies recurrent neural networks to problems whose input and output sequences have different lengths with complicated and non-monotonic relationships (Sutskever, I., Vinyals, O., and Le, Q. V., 2014). Specifically, standard Long Short-Term Memory (LSTM) networks are used due to their superior performance for small corpora, as is the case in TLP, instead of more popular models based on Transformers (Ezen-Can, A., 2020). Also, model awareness of the context (e.g., through the WE) helps understand the semantic meaning of an input sequence and generate a more informative response (Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., and Jiang, M., 2020). Considering all these points, Figure 13 shows the diagram of the proposed causality-contextualized S2S language model using the LSTM and the CCWE. Note that given the sequential nature of S2S, the input/output interface to the system is no longer a BOW but a one-hot encoded single-word vector, i.e., words are presented and retrieved from the language model on a one-by-one basis.

Table 7 shows the plain Root Cause outputs obtained from the system given potential generic Problems. In light of these results, the causality-contextualized language model exhibits the performance of a "pidgin", and this is mainly attributed to the strict lexicon-driven text preprocessing stage. The model does not retrieve the ROX entries literally. Instead, it displays a generalization capacity using vague words (e.g., most Problems are blamed on "reporting" as their Root Cause). Such pathological utterances, also known as *hallucinations*, are common with S2S (Lee, K., Firat, O., Agarwal, A., Fannjiang, C., and Sussillo, D., 2018). And due to the discrepancy between this vaguely generated text and the detailed ROX reports, the exposure bias problem that usually affects such autoregressive language models is increasingly more penalizing for technical language (Wang, C., and Sennrich, R., 2020). Also, input Problems need to be provided using long, elaborate and verbose descriptions, otherwise the model outputs nothing (i.e., long chains of padding symbols). This may be attributed to the most critical components of the LSTM cell, i.e., the forget gate and the activation function (Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J., 2017).

Finally, there are diminishing returns with increasing the scale of model parameters, dataset size, and training computation, because these variables are power laws (Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D., 2020), so the potential for significant improvement needs to be driven by a complementary source of knowledge, such as the FMMEA, as it has been researched in this work. The causal structure of use here shows 19 Failure Mechanisms for 12 Failure Modes regarding 14 components, so further refinements (or generalizations) may be observed if these values are augmented.

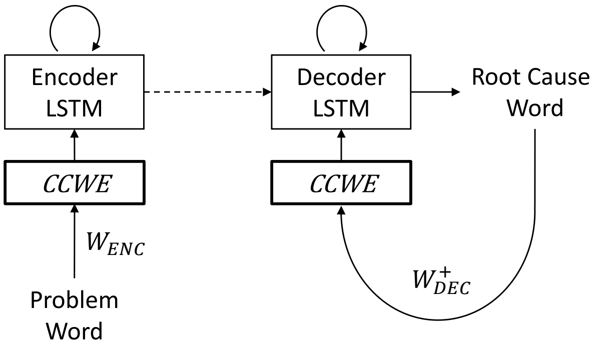

Figure 13. Diagram of the causality-contextualized S2S language model using the LSTM and the CCWE.

Table 7. Plain troubleshooting Root Cause sentences generated by the causality-contextualized language model given potential generic Problems.

Problem	Root Cause
oil leak found on bogie, gear box, and wheel at high speed	report design
hot axle box bearing	assembly
traction motor caught fire, smoke alert on commercial service	report inspection
noisy blower does not turn: power electronics are not available	report failure part

5. CONCLUSION

This work describes a first exploratory work on how the Causal Inference paradigm may be applied to troubleshooting rolling stock bogies through the extraction of linguistic knowledge from FMMEA and ROX text data using graphs and contextualized word embeddings. The overall conclusions indicate that the inference of causality has already been attained with the available theoretical and practical documentation, showing a consensus greater than 70%. Interestingly, though, some disagreement between Root Cause and Problem has arisen in a few areas, leading to poor diagnosis results, and potentially indicating that textual expression improvements are necessary in the technical materials.

The central piece of this research is the construction of a neural word embedding that differs from the state of the art, which is focused on modeling the local context of a single word. The proposed model jointly embeds two whole textual instances that belong to different (causal) contexts. In terms of evaluation, given that CI is not a well-defined task in language processing, the results may be questioned due to their strict dependence on subjective human criteria. This is a clear point of general improvement (beyond the specific purposes of this work) toward the fair assessment of other related CI approaches such as the Twin Networks method to estimate the probabilities of causation (Vlontzos, A., Kainz, B., and Gilligan-Lee, C. M., 2021), the causal regularization of neural networks to improve their interpretability (Bahadori, M. T., Chalupka, K., Choi, E., Chen, R., Stewart, W. F., and Sun, J., 2017; Shen, Z., Cui, P., Kuang, K., Li, B., and Chen, P., 2018), or the learning of causally disentangled representations using Variational Autoencoders (Suter, R., Miladinović, D., Schölkopf, B., and Bauer, S.,, 2019; Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J., 2020).

In terms of application, a direct implementation of this developing approach could be driven by a retrieval-augmented generation system for work orders to advise the maintenance team by identifying the most probable underlying root cause to a given problem, and reduce both the time to action and asset downtime while increasing the safety of the railway service (Ansaldi, S. M., Agnello, P., Pirone, A., and Vallerotonda, M. R., 2021). This enhanced troubleshooting system would equip a model that combines pre-trained parametric memory (i.e., the causality-contextualized word embedding) and non-parametric memory (i.e., a classic data retrieval-based engine) for language generation (Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t-, Rocktäschel, T., Riedel, S., and Kiela, D., 2020). However, the shortage of maintenance text data may hinder the exploitation of this approach. Therefore, a NLP augmentation strategy could be helpful (Bayer, M., Kaufhold, M.-A., Buchhold, B., Keller, M., Dallmeyer, J., and Reuter, C., 2021), although the larger the data analyzed, the greater the chance that spurious correlations dominate the results and lead to erroneous conclusions (Dima, A., Lukens, S., Hodkiewicz, M., Sexton, T., and Brundage, M. P., 2021). Alternatively, fine-tuning a bigger pre-trained language model, which has become the de facto standard for doing transfer learning in NLP, could also be advantageous (Li, J., Tang, T., Zhao, W. X., and Wen, J.-R., 2021). Finally, the deployment of the presented approach to a different railway PHM asset such as the Passenger Door System may reveal further CI insights into the integration of FMMEA with ROX (Dinmohammadi, F., Alkali, B., Shafiee, M., Bérenguer, C., and Labib, A., 2016), and with the increased availability of diverse SCM, a Graph Neural Network could expect to learn a truly holistic troubleshooting system at the train level (Bronstein, M. M., Bruna, J., Cohen, T., Velickovic, P., 2021).

ACKNOWLEDGMENT

We would like to show our gratitude to Prof. Francesc Alías, Dr. Jonathan Brown, and Dr. Eduardo Di-Santi for their insightful comments which greatly improved the manuscript. The contribution of Alexandre Trilla to this research was partially supported by the Government of Catalonia (Generalitat de Catalunya) Grant No. 2020 DI 54.

REFERENCES

Ahmed, O., Träuble, F., Goyal, A., Neitz, A., Bengio, J., Schölkopf, B., Wüthrich, M., and Bauer, S. (2020). CausalWorld: A Robotic Manipulation Benchmark for Causal Structure and Transfer Learning. *arXiv:2010.04296 [cs.RO]*, 1–18.

Alaa, A. M., Weisz, M., and van der Schaar, M. (2017). Deep Counterfactual Networks with Propensity-Dropout. *Proc. of the 34th International Conference on Machine Learning*, 1–6.

Almeida, F., and Xexéo, G. (2019). Word Embeddings: A Survey. *arXiv:1901.09069 [cs.CL]*, 1–10.

Ansaldi, S. M., Agnello, P., Pirone, A., and Vallerotonda, M. R. (2021). Near Miss Archive: A Challenge to Share Knowledge among Inspectors and Improve Seveso Inspections. *Sustainability*, *13*(8456), 1–21.

Ansari, F. (2020). Cost-based text understanding to improve maintenance knowledge intelligence in manufacturing enterprises. *Computers and Industrial Engineering*, *141*(106319).

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant Risk Minimization. *arXiv:1907.02893 [stat.ML]*, 1–31.

Atamuradov, V., Medjaher, K., Dersin, P., Lamoureux, B., and Zerhouni, N. (2017). Prognostics and Health Management for Maintenance Practitioners - Review, Implementation and Tools Evaluation. *International Journal of Prognostics and Health Management*, *8*(60), 1–31.

Bahadori, M. T., and Heckerman, D. E. (2021). Debiasing Concept Bottleneck Models with a Causal Analysis Technique. *Proc. of the International Conference on Learning Representations*, 1–11.

Bahadori, M. T., Chalupka, K., Choi, E., Chen, R., Stewart, W. F., and Sun, J. (2017). Causal Regularization. *arXiv:1702.02604 [cs.LG]*, 1–18.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *Proc. of the International Conference on Learning Representations*, 1–15.

Bakarov, A. (2018). A Survey of Word Embeddings Evaluation Methods. *arXiv:1801.09536 [cs.CL]*, 1–26.

Barocas, S., Selbst, A. D., and Raghavan, M. (2019). The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons. *arXiv:1912.04930 [cs.CY]*, 1–17.

Bayer, M., Kaufhold, M.-A., Buchhold, B., Keller, M., Dallmeyer, J., and Reuter, C. (2021). Data Augmentation in Natural Language Processing: A Novel Text Generation Approach for Long and Short Text Classifiers. *arXiv:2103.14453 [cs.CL]*, 1–20.

Bengio, Y. (2017). The Consciousness Prior. *arXiv:1709.08568 [cs.LG]*, 1–7.

Bronstein, M. M., Bruna, J., Cohen, T., Velickovic, P. (2021). Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv:2104.13478 [cs.LG]*, 1–156.

Brundage, M. P., Sexton, T., Hodkiewicz, M., Dima, A., and Lukens, S. (2021). Technical language processing: Unlocking maintenance knowledge. *Manufacturing Letters*, *27*, 42–46.

Brundage, M. P., Weiss, B. A., and Pellegrino, J. (2020). Summary Report: Standards Requirements Gathering Workshop for Natural Language Analysis. *National Institute of Standards and Technology Advanced Manufacturing Series*, *100*(30), 1–50.

Cai, H., Zheng, V. H., and Chang, K. C.-C. (2018). A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. *IEEE Transactions on Knowledge and Data Engineering*, *30*, 1616–1637.

Camacho-Collados, J., and Pilehvar, M. T. (2018). From Word to Sense Embeddings: A Survey on Vector Representations of Meaning. *Journal of Artificial Intelligence Research*, *63*, 743–788.

Chen, W., Grangier, D., and Auli, M. (2015). Strategies for Training Large Vocabulary Neural Language Models. *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, 1975–1985.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 1–14.

Conrad, S. (2019). Register in English for Academic Purposes and English for Specific Purposes. *Register Studies*, *1*(1), 168–198.

Crawshaw, M. (2020). Multi-Task Learning with Deep Neural Networks: A Survey. *arXiv:2009.09796 [cs.LG]*, 1–43.

Dima, A., Lukens, S., Hodkiewicz, M., Sexton, T., and Brundage, M. P. (2021). Adapting natural language processing for technical text. *Applied AI Letters*, *2*(e33), 1–11.

Dinmohammadi, F., Alkali, B., Shafiee, M., Bérenguer, C., and Labib, A. (2016). Risk Evaluation of Railway Rolling Stock Failures Using FMECA Technique: A Case Study of Passenger Door System. *Urban Rail Transit*, *2*(3–4), 128–145.

DuBay, W. H. (2004). The Principles of Readability. *Impact Information*, 1–77.

Ebrahimipour, V., Rezaie, K., and Shokravi, S. (2010). An ontology approach to support FMEA studies. *Expert Systems with Applications*, *37*(1), 671–677.

Ezen-Can, A. (2020). A Comparison of LSTM and BERT for Small Corpus. *arXiv:2009.05451 [cs.CL]*, 1–12.

Fink, O., Wang, Q., Svensén, M., Dersin, P., Lee, W.-J., and Ducoffe, M. (2020). Potential, challenges and future directions for deep learning in prognostics and health management applications. *Engineering Applications of Artificial Intelligence*, *92*(103678), 1–15.

Gelman, A., and Imbens, G. (2013). *Why ask why? Forward causal inference and reverse causal questions* (Tech. Rep. No. 19614). National Bureau of Economic Research.

Gelman, A., and Vehtari, A. (2020). What are the most important statistical ideas of the past 50 years? *arXiv:2012.00174 [stat.ME]*, 1–19.

Goth, G. (2016). Deep or Shallow, NLP Is Breaking Out. *Communications of the ACM*, *59*(3), 13–16.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222–2232.

Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. (2020). A Survey of Learning Causality with Data: Problems and Methods. *ACM Computing Surveys*, *53*(4).

Gutmann, M., and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *Proc. of the 13th International Conference on Artificial Intelligence and Statistics*, 297–304.

Hancock, J. T., and Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, *7*(28), 1–41.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A Flexible Approach for Counterfactual Prediction. *Proc. of the 34th International Conference on Machine Learning*, 1–10.

Hastings, E. M., Sexton, T., Brundage, M. P., and Hodkiewicz, M. (2019). Agreement Behavior of Isolated Annotators for Maintenance Work-Order Data Mining. *Proc. of the Annual Conference of the Prognostics and Health Management Society*, 1–7.

Imbens, G. W. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *arXiv:1907.07271 [stat.ME]*, 1–76.

ISO. (2003). *Condition monitoring and diagnostics of machines – General guidelines on data interpretation and diagnostics techniques* (Tech. Rep. No. 13379:2003(E)). International Organization for Standardization.

Iyyer, M., Boyd-Graber, J., Claudino, L., Socher, R., and

Daumé III, H. (2014). A Neural Network for Factoid Question Answering over Paragraphs. *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 633–644.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. (2016). Exploring the Limits of Language Modeling. *arXiv:1602.02410 [cs.CL]*, 1–11.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling Laws for Neural Language Models. *arXiv:2001.08361 [cs.LG]*, 1–30.

Karray, M. H., Ameri, F., Hodkiewicz, M., and Louge, T. (2019). ROMAIN: Towards a BFO compliant reference ontology for industrial maintenance. *Applied Ontology*, *14*(2), 155–177.

Le, Q., and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proc. of the 31st International Conference on Machine Learning*, 1–9.

Leao, B. P., Fitzgibbon, K. T., Puttini, L. C., and de Melo, G. P. B. (2008). Cost-Benefit Analysis Methodology for PHM Applied to Legacy Commercial Aircraft. *Proc. of IEEE Aerospace Conference*, 1–13.

LeCun, Y. and Bengio, Y., and Hinton, G. E. (2015). Deep Learning. *Nature*, *521*, 436-444.

Lee, K., Firat, O., Agarwal, A., Fannjiang, C., and Sussillo, D. (2018). Hallucinations in Neural Machine Translation. *Proc. of the 32th Conference on Neural Information Processing Systems*, 1–18.

Leuenberger, H., Puchkov, M., and Schneider, B. (2013). Right, First Time Concept and Workflow. A Paradigm Shift for a Smart & Lean Six-sigma Development. *Swiss Pharma*, *35*(3), 3–16.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, *2*, 302–308.

Levy, O., and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Proc. of the 27th International Conference on Neural Information Processing Systems*, *2*, 2177–2185.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t-, Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Proc. of the International Conference on Neural Information Processing Systems*, 1–18.

Li, J., Tang, T., Zhao, W. X., and Wen, J.-R. (2021). Pretrained Language Models for Text Generation: A Survey. *arXiv:2105.10311 [cs.CL]*, 1–9.

Li Y., and Yang T. (2018). Word Embedding for Understanding Natural Language: A Survey. *Guide to Big Data Applications. Studies in Big Data*, *26*, 83–104.

Li, Y.-H., Wang, Y.-D., and Zhao, W.-Z. (2009). Bogie Failure Mode Analysis for Railway Freight Car Based on FMECA. *Proc. of the 8th International Conference on Reliability, Maintainability and Safety*, 5–8.

Liu, Q., Kusner, M. J., and Blunsom, P. (2020). A Survey on Contextual Embeddings. *arXiv:2003.07278 [cs.CL]*, 1–13.

Liu, Z., Jia, Z., Vong, C.-M., Han, W., Yan, C., and Pecht, M. (2018). A Patent Analysis of Prognostics and Health Management (PHM) Innovations for Electrical Systems. *IEEE Access*, *6*, 18088–18107.

Maguire, P., Mulhall, O., Maguire, R., and Taylor, J. (2015). Compressionism: A Theory of Mind Based on Data Compression. *Proc. of the 11th International Conference on Cognitive Science*, 294–299.

Mathew, S., Das, D., Rossenberger, R., and Pecht, M. (2008). Failure Mechanisms Based Prognostics. *Proc. of the International Conference on Prognostics and Health Management*, 1–6.

Mihalcea, R., and Radev, D. (2011). *Graph-Based Natural Language Processing and Information Retrieval*. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proc. of Workshop at the International Conference on Learning Representations*, 1–12.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Proc. of the Conference on Neural Information Processing Systems*, 1–9.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *Proc. of the North American Chapter of the Association for Computational Linguistics*, 746–751.

Mitrovic, J., McWilliams, B., Walker, J., Buesing, L., and Blundell, C. (2021). Representation Learning via Invariant Causal Mechanisms. *Proc. of the International Conference on Learning Representations*, 1–12.

Mnih, A., and Teh, Y. W. (2012). A Fast and Simple Algorithm for Training Neural Probabilistic Language Models. *Proc. of the 29th International Conference on Machine Learning*, 1–8.

Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. *Proc. of the Conference on Fairness, Accountability, and Transparency*, 1–13.

Nastase, V., Mihalcea, R., and Radev, D. (2015). A survey of graphs in natural language processing. *Natural Language Engineering*, *21*(5), 665–697.

Naveed, A., Li, J., Saha, B., Saxena, A., and Vachtsevanos, G. (2012). A Reasoning Architecture for Expert Troubleshooting of Complex Processes. *Proc. of the Annual Conference of the Prognostics and Health Management*

Society, 1–8.

Navinchandran, M., Sharp, M. E., Brundage, M. P., and Sexton, T. B. (2019). Studies to Predict Maintenance Time Duration and Important Factors From Maintenance Workorder Data. *Proc. of the Annual Conference of the Prognostics and Health Management Society*, 1–11.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54–60.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv:1802.05365 [cs.CL]*, 1–15.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S. (2019). Language Models as Knowledge Bases? *Proc. of the Conference on Empirical Methods in Natural Language Processing*, 1–11.

Polenghi, A., Roda, I., Macchi, M., and Pozzetti, A. (2022). Ontology-augmented Prognostics and Health Management for shopfloor-synchronised joint maintenance and production management decisions. *Journal of Industrial Information Integration*, *27*(100286).

PwC. (2019). *It's time for a consumer-centred metric: introducing 'return on experience'. Global Consumer Insights Survey* (Tech. Rep. No. 512587-2019). PricewaterhouseCoopers International Limited.

Rehman, Z., and Kifor, C. V. (2016). An Ontology to Support Semantic Management of FMEA Knowledge. *International Journal of Computers Communications & Control*, *11*(4), 507–521.

Salakhutdinov, R., and Hinton, G. (2009). Semantic hashing. *International Journal of Approximate Reasoning*, *50*, 969–978.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Towards Causal Representation Learning. *Proc. of the IEEE*, *109*(5), 612–634.

Sexton, T. B., and Brundage, M. P. (2019). Nestor: A Tool for Natural Language Annotation of Short Texts. *Journal of Research of National Institute of Standards and Technology*, *124*(124029), 1–5.

Sexton, T., Hodkiewicz, M., Brundage, M. P., and Smoker, T. (2018). Benchmarking for Keyword Extraction Methodologies in Maintenance Work Orders. *Proc. of the Annual Conference of the Prognostics and Health Management Society*, 1–10.

Shalit, U., Johansson, F. D., and Sontag, D. (2016). Estimating individual treatment effect: generalization bounds and algorithms. *arXiv:1606.03976 [stat.ML]*, 1–20.

Sharma, A., and Kiciman, E. (2020). DoWhy: An End-to-End Library for Causal Inference. *arXiv:2011.04216 [stat.ME]*, 1–5.

Sharp, M. E., Sexton, T. B., and Brundage, M. P. (2017). Semi-Autonomous Labeling of Unstructured Maintenance Log Data for Diagnostic Root Cause Analysis. *Proc. of the International Conference Advances in Production Management Systems*, 1–8.

Shen, Z., Cui, P., Kuang, K., Li, B., and Chen, P. (2018). Causally Regularized Learning with Agnostic Data Selection Bias. *Proc. of ACM Multimedia Conference*, 1–9.

Smetkowska, M., and Mrugalska, B. (2018). Using Six Sigma DMAIC to Improve the Quality of the Production Process: A Case Study. *Procedia – Social and Behavioral Sciences*, *238*, 590–596.

Stampe, D. W. (2008). Towards A Causal Theory of Linguistic Representation. *Midwest Studies in Philosophy*, *2*(1), 42–63.

Su, Y., Awadallah, A. H., Wang, M., and White, R. H. (2018). Natural Language Interfaces with Fine-Grained User Interaction: A Case Study on Web APIs. *Proc. of the 41th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1–10.

Suter, R., Miladinović, D., Schölkopf, B., and Bauer, S.,. (2019). Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness. *Proc. of the 36th International Conference on Machine Learning*, 1–10.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *arXiv:1409.3215 [cs.CL]*, 1–9.

Tan, L., Zhang, H., Clarke, C. L. A., and Smucker, M. D. (2015). Lexical Comparison Between Wikipedia and Twitter Corpora by Using Word Embeddings. *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, 657–661.

Tan, S., Zhou, Z., Xu, Z., and Li, P. (2019). On Efficient Retrieval of Top Similarity Vectors. *Proc. of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 5236–5246.

Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavaf, N., and Fox, E. A. (2020). Natural Language Processing Advancements By Deep Learning: A Survey. *arXiv:2003.01200 [cs.CL]*, 1–21.

Trilla, A., Bob-Manuel, J., Lamoureux, B., and Vilasis-Cardona, X. (2021). Integrated Multiple-Defect Detection and Evaluation of Rail Wheel Tread Images using Convolutional Neural Networks. *International Journal of Prognostics and Health Management*, *12*(1), 1–19.

Vlontzos, A., Kainz, B., and Gilligan-Lee, C. M. (2021).

Estimating the probabilities of causation via deep monotonic twin networks. *arXiv:2109.01904 [cs.LG]*, 1–10.

Wang, C., and Sennrich, R. (2020). On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation. *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, 3544–3552.

Wang, Z., and Wang, H. (2016). Understanding Short Texts. *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*, 1–4.

Yaghoobzadeh, Y., Kann, K., Hazen, T. J., Agirre, E., and Schütze, H. (2019). Probing for Semantic Classes: Diagnosing the Meaning Content of Word Embeddings. *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, 5740–5753.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. (2020). CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. *arXiv:2004.08697 [cs.LG]*, 1–21.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. (2020). A Survey on Causal Inference. *arXiv:2002.02770 [stat.ME]*, 1–38.

Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., and Jiang, M. (2020). A Survey of Knowledge-Enhanced Text Generation. *arXiv:2010.04389 [cs.CL]*, 1–44.

Zanette, D. H., and Montemurro, M. A. (2005). Dynamics of Text Generation with Realistic Zipf's Distribution. *Journal of Quantitative Linguistics*, *12*(1), 29–40.

Zheng, M., Marsh, J. K., Nickerson, J. V., and Kleinberg, S. (2020). How causal information affects decisions. *Cognitive Research: Principles and Implications*, *5*(6), 1–24.

BIOGRAPHIES

Alexandre Trilla graduated from La Salle University of Barcelona with a M.Sc. in Electronics and Telecommunications Engineering in 2008, and a M.Sc. in IT Management in 2010. He has an academic research background in spoken language processing, and an industrial research background in PHM. He has authored several publications in scientific conferences and journals (International Journal of Prognostics and Health Management, IEEE Transactions on Audio, Speech, and Language Processing, Chemical Engineering Transactions, and the Journal of Rail and Rapid Transit). At present, he is a Senior Data Scientist and R&D Program Manager at Alstom, working on the deployment of PHM to the railway environment. He leads the development of predictive maintenance based on Machine Learning, and he is especially interested in building solutions using artificial neural networks and Deep Learning.

Nenad Mijatovic is a Data Science Leader in Alstom. He has over 20 years of algorithm development experience in a variety of areas, such as statistics, numerical optimization, machine learning, AI, and causality. Before joining Alstom, Dr. Mijatovic has held several R&D and leadership positions in the industry, from startups to blue-chip companies. His interests are applying machine learning and AI methods for industrial applications. In his current position, Dr. Mijatovic leads Alstom's data science teams responsible for delivering industrial-grade ML and AI algorithms for maintenance, operations, energy, and city flow solutions.

Xavier Vilasis-Cardona is full professor at La Salle, Universitat Ramon Llull, Barcelona. He holds a degree in physics ('89) and a PhD in physics ('93) by Universitat de Barcelona. He is member of the IEEE, of the IEEE CNNAC technical committee and of the LHCb collaboration. He is currently leading the Data Science for the Digital Society (DS4DS) research group.