# Preliminary Study of Automated Analysis of Nuclear Power Plant Event Reports Based on Natural Language Processing Techniques

**Yunfei Zhao[a*], Xiaoxu Diao[a], and Carol Smidts[a]**
[a] Nuclear Engineering Program, Department of Mechanical and Aerospace Engineering, The Ohio State University, Columbus, USA

**Abstract:** A large number of licensee generated event reports are available in the nuclear power generation sector. A comprehensive analysis of the reports may provide valuable insights for improving nuclear power plant operation and safety. However, the free text format of the reports poses great challenges to analysis. To address this issue, we propose an automated analysis approach based on natural language processing techniques. The preliminary study in the paper is carried out for the U.S. Nuclear Regulatory Commission Licensee Event Report database with the main objective of identifying the causal relationships between events described in a report. To this end, a list of keywords (e.g., *caused*, *due to*) that indicate causal relationships are first identified based on a thorough review of a set of sample reports. These keywords are then considered together with part of speech tagging and dependency parsing of a sentence, which are implemented using the Stanford CoreNLP API, to identify the causal and consequent events in the sentence. The list of keywords and the analysis results for two of these keywords, *caused* and *caused by*, are presented. The research investigates the feasibility of and provides the basis for developing an automated tool for analyzing text-based event reports.

**Keywords:** Natural Language Processing, Text-Based Report, Causal Relationship, Licensee Event Report, Nuclear Power Plant.

## 1. INTRODUCTION

The long operating history of nuclear power plants has enabled regulatory agencies to collect a large volume of data relating to plant operating experience. Examples include the Licensee Event Report (LER) database maintained by the U.S. Nuclear Regulatory Commission (NRC) [1], the Institute of Nuclear Power Operation (INPO) Operating Experience database [2], and the Chinese Operating Experience Feedback System [3]. The reports in these databases are a source of valuable information on plant operations. A comprehensive analysis of the reports may provide insights for improving plant operation and safety, for instance preventing recurrence of similar events [4] and taking proactive mitigation measures before serious consequences materialize.

Because of the potential benefits of report analysis and advancements in natural language processing (NLP) and data mining, we have seen an increasing use of text-based reports to improve system performance. Within the nuclear industry, the relative importance of significant performance shaping factors (PSFs) was extracted from event investigation reports of domestic nuclear power plants [5]. In this analysis, the events in a report were identified and marked corresponding to predefined PSFs. In [3], correlation analysis, cluster analysis, and association rule mining were implemented to identify the intrinsic correlations among causal factors from human factor event reports, where causal factors applicable to each report were identified before the analyses. In [4], a method for ranking groups of similar events based on pre-determined indexes was presented and applied to reports from the LER database. The results were used to support the prioritization of further investigation. In [6], a big data-theoretic approach for the quantification of organizational failure mechanisms was proposed, where written documents and texts serve as the sources of data, though the method used for text mining was not presented. Outside the nuclear industry, [7] proposed a method for estimating system reliability based on text form records, and applied the method to two examples, a Wheatstone bridge and a safety

[*] Correspondence to Yunfei Zhao at: zhao.2263@osu.edu

instrumentation system. In the method, the content of the text was classified into predefined classes (e.g., *good*, *bad*). In [8], a semi-automated technique for classifying text-based close call reports from the railway industry was introduced. The classification schema used natural language processing techniques to classify the reports in accordance with the threat pathways shown on bow-tie diagrams. In [9], an approach was developed to identify high-risk sequences of events in general aviation accidents based on historical accident reports, which were generated using well defined codes.

Although there have been significant advancements in using event reports to gain insights in various industries, no research has, to the authors' best knowledge, been done toward developing event scenarios based on automated analysis of reports. This research is needed to accelerate the analysis of reports and to aid the analysts in understanding the events better. In addition, it provides the basis for statistical analysis of the events to gain further insights. However, most reports are written in unstructured free text format, which poses great challenges to analysis. To address this issue, we propose an automated analysis approach based on natural language processing techniques. The preliminary study is carried out for the LER database with the objective of extracting the event scenario from the unstructured free text in a report. The method will be extended to other sources of data related to nuclear power plant events, such as U.S. NRC Augmented Inspection Team (AIT) reports and the NRC Office for Analysis and Evaluation of Operational Data, in future research.

The rest of the paper is organized as follows. Section 2 provides a brief introduction to the U.S. NRC LER database. In Section 3, the proposed method is introduced and the results of analysis are presented. Section 4 concludes the paper with a summary of the research and discussions on future research.

## 2. LER STRUCTURE

Commercial nuclear reactor licensees in the United States are required to report certain event information per 10CFR50.73 [10]. The event reports constitute the LER database. The database allows users to search for reports based on a variety of criteria, including title, dates, plant characteristics, event characteristics, and abstract or document keywords. As of March 20, 2018, the database has collected 53,016 event reports, starting from 1980. At the beginning of each report is a structured form consisting of 15 required fields, such as facility name, report title, event date, and report date. Following the form is the main section of the report, which is comprised of an abstract and a narrative of the event. The narrative of each report usually presents the plant conditions before the event, a description of the event, significant safety consequences and implications, the cause of the event, corrective actions taken, and previous occurrences, if applicable. The abstract summaries the progression of the event, the consequences, and the corrective actions. Both the abstracts and narratives are written in free text, which provides flexibility of reporting, but impedes the analysis of the reports. An example abstract is given below for illustration:

"*On January 11, 2018, at 10:41 EST, a planned train swap of the Reactor Building Heating Ventilation and Air Conditioning (RB 1VAC) system resulted in the Technical Specification ('1'S) for secondary containment (SC) pressure boundary not being met for less than one minute. The maximum secondary containment pressure recorded during that time was approximately 0.116 inches of vacuum water gauge. Secondary containment pressure was restored to within TS limits by starting Division 1 of the Standby Gas Treatment System (SLITS). There were no safety consequences or radiological releases associated with this event. The cause of this momentary loss of SC was determined to be the effect of the RBHVAC West Exhaust Fan Modulating Damper failing to fully open. For corrective actions, [the plant] has repaired the damper.*"

A preliminary review of the reports found that the most important information pertaining to an event can be found in the title and abstract. As a result, the analysis in this paper will be limited to identifying the causal relationships between events described in the title and abstract of a report. However, the proposed method can be adapted to the narrative section with little effort. In future research, the difference between analyses with and without the narrative section will be compared.

Besides, the method for handling the structured information in the reports will be investigated in future research.

## 3. ANALYSIS BASED ON NLP TECHNIQUES

### 3.1. Overview

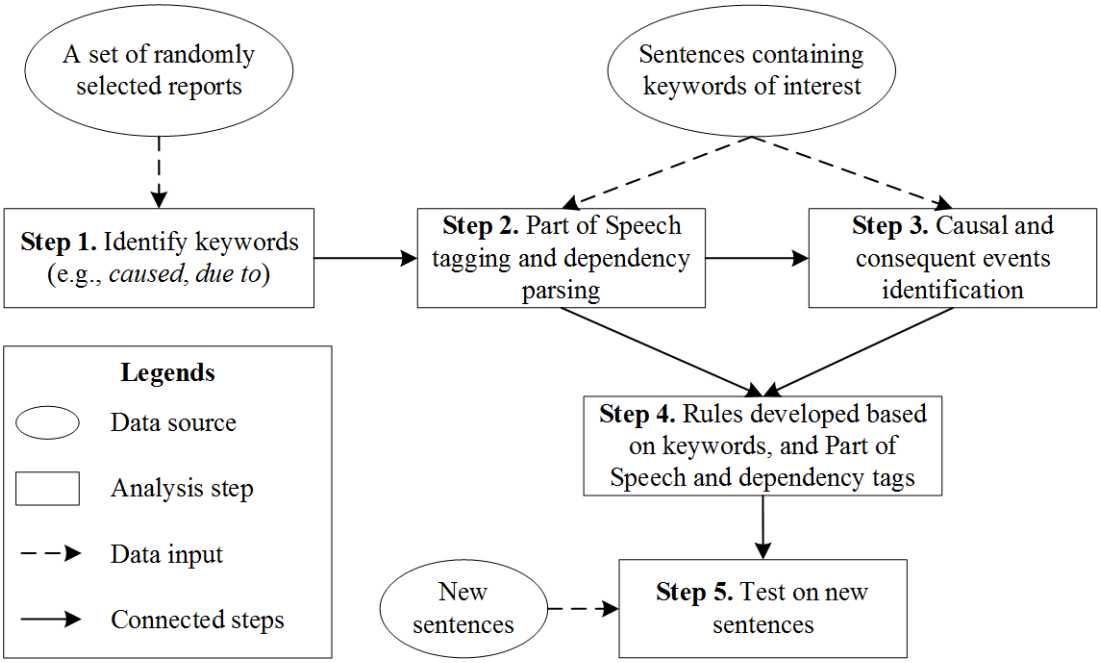


**Figure 1. Overview.**

An overview of the analysis is shown in Figure 1. It consists of five steps. In the first step, a list of keywords indicating causal relationships is developed from a set of randomly selected reports. To enable automated identification of the causal and consequent events, a set of rules are developed based on Part of Speech tagging, dependency parsing, and manual analysis of the causal and consequent events in sentences containing keywords of interest. In the fifth step, the developed rules are tested on new sentences to assess the effectiveness of the rules in identifying causal relationships from free text. Each step is introduced in further detail in the following sections.

### 3.2. Keywords Identification

Manual analysis of about 20 reports led to the observation that in most cases the causal relationships between events can be pinpointed by certain keywords. Examples of these keywords include *resulted in* and *the cause of* in the example abstract in Section 2, as well as *caused*, *caused by*, *due to*, *because of*, etc. Apparently, these keywords typically indicate causal relationships. Based on this observation, it was determined that a relatively small set of keywords could cover most causal relationships in the reports. So one could start by developing the set of keywords and use it to help analyze the reports. To investigate this idea, we randomly selected 70 reports from the LER database, and analyzed the title and abstract of each report sequentially. While doing this, we recorded the newly discovered keywords from each consecutive report analyzed. After completing the analysis for the 70 reports, the relationships between the total and incremental numbers of keywords and the sequential report number were plotted, as shown in Figure 2, where it can be seen that the number of new keywords decreases rapidly and the total number of keywords increases very slowly after analyzing about 30 reports. It needs to be noted that the analysis was limited to the title and abstract of the reports. If the analysis is extended to the narrative section, which contains more content, the trend will be even more apparent. It also needs to be noted that in certain cases the causal relationships cannot be captured simply by these keywords, such as the causal relationship implied by time sequence. However, these cases are

few. Resolving these cases is one of the focuses of future research. From the 70 reports, we identified 24 keywords, which are listed in Table 1. In the subsequent analysis in this paper, we will focus on two keywords: *caused* and *caused by*. The analysis for the other keywords follows the same procedure.

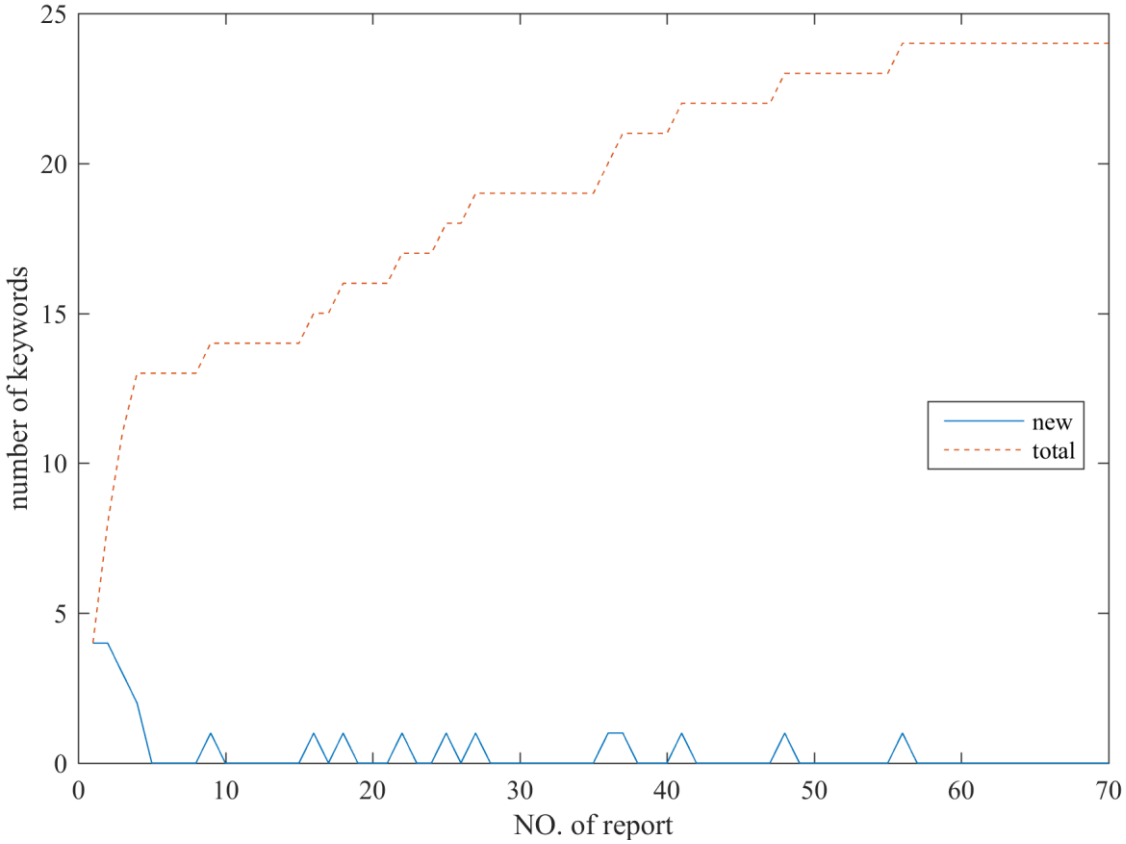


**Figure 2. Keywords Identification.**

**Table 1. Keyword List.**

| NO. | Keyword | NO. | Keyword |
|---|---|---|---|
| 1 | *Result in* | 13 | *As a result* |
| 2 | *Caused* | 14 | *Because* |
| 3 | *Due to* | 15 | *Impact* |
| 4 | *Be caused by* | 16 | *Be attributed to* |
| 5 | *Result from* | 17 | *Initiate* |
| 6 | *Follow* | 18 | *Produce* |
| 7 | *As a result of* | 19 | *Lead to* |
| 8 | *[the/a ...] cause(s) of* | 20 | *Actuate* |
| 9 | *Be followed by* | 21 | *Because of* |
| 10 | *Be due to* | 22 | *Be required to* |
| 11 | *[the/a ...] cause(s) be* | 23 | *Contributing factor(s) be* |
| 12 | *In response to* | 24 | *Give rise to* |

### 3.3. Part of Speech Tagging and Dependency Parsing

There are, of course, innumerable variations of word combinations possible in a given sentence. So even if the analysis is constrained to a limited number of keywords, it is still impossible to directly develop the rules for identifying causal relationships between events. In this paper, we use the part of speech for each word and the dependencies between the keyword and other words in a sentence as features to develop the rules.

Parts of speech denote word classes or categories, for instance *noun* or *verb*. Parts of speech are useful because of the large amount of information they give about a word and its neighbors [11]. A number of lists of parts of speech have been defined by different researchers. The most widely used list is the 45-tag Penn Treebank tagset [12]. Dependency parsing is used to identify the semantic relations between words in a sentence. Examples of dependencies include nominal subject (*nsubj*), direct object (*dobj*), determiner (*det*), etc [11]. For instance, in the sentence "She looks very beautiful.", *She* is the nominal subject (*nsubj*) of *looks*. Further details on dependency relations can be found in [13,14]. Both parts of speech and dependencies have been used as features in more complex natural language processing tasks, such as semantic role labelling [15].

In this paper, the Stanford CoreNLP API [16] is used to implement Part of Speech tagging and dependency parsing. The application of this API for these two tasks is demonstrated in Figure 3, where Part of Speech tags are placed on the top of each word and dependencies are represented as directed, labeled arcs.

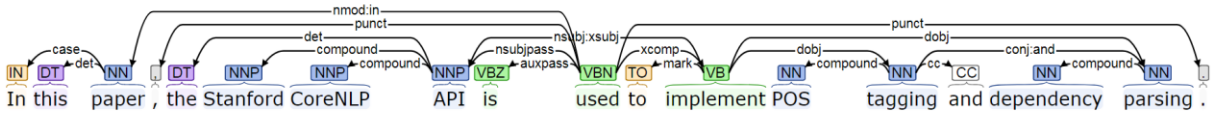


**Figure 3. Illustration of Applying the Stanford CoreNLP API in Part of Speech Tagging and Dependency Parsing.**

### 3.4. Causal and Consequent Events Identification

The objective of the analysis in this section is to identify the causal and consequent events linked by the keywords of interest, *caused* and *caused by*, based on the part of speech of each word and dependencies between the keywords and other words in a sentence. It needs to be noted that an event is usually described by a phrase which is comprised by several words rather than one. As a preliminary study, we only focus on the single word that represents the causal event or consequent event in this paper. Expanding the analysis to capture the phrase that actually represents the event is the content of future research. Thirty sentences were analyzed for *caused* and twenty-six sentences were analyzed for *caused by*. Two examples of the analysis for *caused* and *caused by* are shown in Figure 4 and Figure 5, respectively. In the first example, the causal relationship pertaining to *caused* is identified as "*failure of transistor Q2*" caused "*the power supply output to go to its maximum value*". In the analysis, the causal event is determined to be *Q2*, and the consequent event is determined to be *output*. The part of speech for both of them is common noun (*NN*), and the dependencies between the two words and the keyword *caused* are nominal subject (*nsubj*) and direct object (*dobj*), respectively. In the second example, the causal relationship pertaining to *caused by* is identified as "*a failed exciter rectifier*" caused "*a loss of generator field excitation*". The single words representing the causal and consequent events are identified as *rectifier* and *excitation*, respectively. The part of speech for both of them is proper noun (*NNP*), and the dependencies between the two words and the keyword *caused by* are nominal modifier with by (*nmod:by*) and clausal modifier of noun (*acl*), respectively.

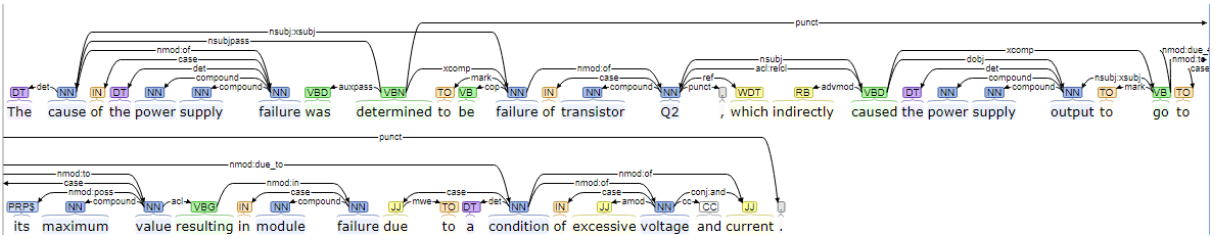


**Figure 4. Example of Causal and Consequent Events Identification for Keyword *Caused*.**

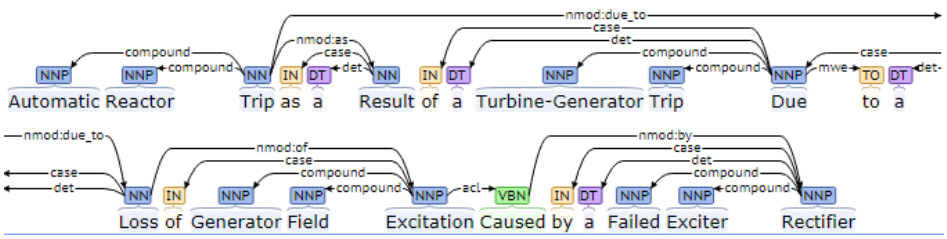


**Figure 5. Example of Causal and Consequent Events Identification for Keyword *Caused by*.**

### 3.5. Rules Development

Based on the results of the manual analysis of 56 sentences to identify causal and consequent events, patterns can be summarized and used to develop the rules for automated analysis. The developed rules for *caused* and *caused by* are listed in Table 2 and Table 3, respectively. The examples in Figure 4 and Figure 5 correspond to the third rule in Table 2 and the second rule in Table 3, respectively.

**Table 2. Rules for *Caused*.**

| NO. | Causal Event | | Consequent Event | |
|---|---|---|---|---|
| | Dependency | Part of Speech | Dependency | Part of Speech |
| 1 | *nsubj* | *NN* | *dobj* | *NNS* |
| 2 | *nsubj* | *NNS* | *dobj* | *JJ* |
| 3 | *nsubj* | *NN* | *dobj* | *NN* |
| 4 | *nsubj* | *DT* | *dobj* | *NN* |
| 5 | *nsubj* | *VB* | *dobj* | *NN* |
| 6 | *nsubj* | *DT* | *ccomp* | *NN* |
| 7 | *nsubj* | *NN* | *dobj* | *NNP* |
| 8 | *acl:relcl* | *WDT* | *dobj* | *NN* |
| 9 | *nsubj:xsubj* | *NN* | *dobj* | *NN* |

**Table 3. Rules for *Caused by*.**

| NO. | Causal Event | | Consequent Event | |
|---|---|---|---|---|
| | Dependency | Part of Speech | Dependency | Part of Speech |
| 1 | *nmod:by* | *NN* | *acl* | *NN* |
| 2 | *nmod:by* | *NNP* | *acl* | *NNP* |
| 3 | *nmod:agent* | *NN* | *nsubjpass* | *NN* |
| 4 | *nmod:agent* | *NN* | *nsubjpass* | *WDT* |
| 5 | *nmod:agent* | *NN* | *nsubjpass:xsubj* | *NN* |
| 6 | *nmod:agent* | *NNS* | *nsubjpass* | *NNS* |
| 7 | *nmod:by* | *NN* | *acl* | *NNS* |
| 8 | *nmod:agent* | *NN* | *nsubjpass* | *NNS* |
| 9 | *nmod:agent* | *NNP* | *nsubjpass* | *NN* |

Automated analysis of the causal relationships between events using the developed rules can be implemented as follows. Given a sentence from a free text report, it is first checked for the keywords developed in Section 3.2. If there is one or more keywords in the sentence, it means that there is at least one causal relationship in the sentence and it needs to be analyzed in further steps. Otherwise, the sentence is ignored. Then part of speech tagging and dependency parsing of the sentence are implemented using the Stanford CoreNLP API. The words in the sentence that have dependencies with the keyword, and the parts of speech of these words, are compared to the developed rules for the corresponding keyword. In case one rule is matched, the causal event and the consequent event can be identified. If there are more than one keywords in a sentence, the sentence can be analyzed sequentially for each keyword.

The implementation is illustrated with another example in Figure 6. In the first step, the keyword *caused* is found in the sentence. Then parts of speech are tagged and dependencies are determined, as shown in Figure 6. Four dependencies (i.e., *nsubj*, *punct*, *nmod:to*, and *dobj*) exist between *caused* and the other words (i.e., *failure*, ., *fault*, and *conductor*), and the parts of speech of these words are *NN*, ., *NN*, and *NN*, respectively. A comparison between this information and the rules for the keyword *caused* finds that the information matches the third rule in Table 2. Then the causal event and consequent event in this sentence can be identified as *failure* and *conductor*, respectively.

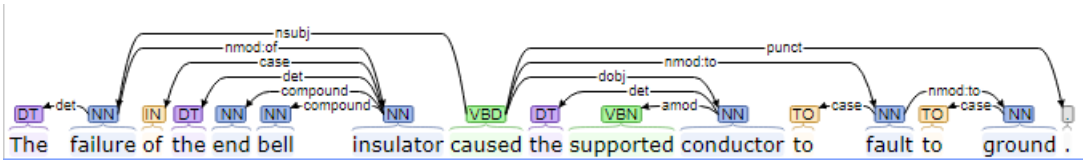


**Figure 6. Illustration of Using Rules for Automated Analysis.**

### 3.6. Test on New Sentences

The feasibility of the proposed approach is investigated by the analysis of new sentences. Twenty sentences containing *caused* and twenty sentences containing *caused by* are randomly selected from reports that have not been investigated in the previous steps. Part of speech tagging and dependency parsing are handled with the Stanford CoreNLP API for each sentence. Then the causal event and consequent event in each sentence are identified and compared with the result based on the developed rules in Table 2 and Table 3. The number of sentences for which the causal relationships can be covered by each rule is summarized in Table 4 and Table 5, for *caused* and *caused by* respectively. The overall coverage of the causal relationships by the rules is 85% for both *caused* and *caused by*. This level of coverage is high for a preliminary study and improving the rules by analyzing more sentences will only increase the performance of the proposed approach.

**Table 4. Test Result for *Caused*.**

| Number of sentences | Corresponding rule NO. |
|---|---|
| 15 | 3 |
| 2 | 1 |
| 3 | No rule matches |

**Table 5. Test Result for *Caused by*.**

| Number of sentences | Corresponding rule NO. |
|---|---|
| 11 | 3 |
| 5 | 1 |
| 1 | 8 |
| 3 | No rule matches |

## 4. CONCLUSIONS AND FUTURE RESEARCH

In this paper, we present a preliminary study of automated analysis of event reports from nuclear power plants. The approach formulated in the paper is based on natural language processing techniques and can be used to identify the causal relationships between events from free text. Two major conclusions can be drawn based on the results. First, the analysis of the reports can start with identifying keywords that indicate causal relationships. It is found in this paper that the set of keywords is of medium size. Second, using parts of speech and dependencies as features in the analysis is feasible and can achieve satisfactory performance. The preliminary study in this paper lays the basis for further research on developing an automated method for free text event report analysis.

It is worth mentioning that identifying causal relationships between events from text-based reports is a challenging project, and further efforts are needed in the future to develop a fully automated method.

Prospective research is expected in areas including: 1) Recognition of patterns pertaining to causal and consequent events based on state-of-the-art machine learning techniques. Neural networks and deep learning techniques have been applied to other tasks in natural language processing, such as semantic role labeling [15] and machine translation [17–19]. These tasks are similar to the task in this paper to a certain extent, and it is believed that the advanced techniques will improve the performance of the proposed approach significantly. 2) Identification of the actual events. The proposed approach in this paper is only able to identify the single word that represents the actual event. How to identify the actual event still needs to be researched. 3) Same events identification. The results of the analysis in this paper are a number of paired events. How to identify the same events among the event repository is critical to developing the entire scenario of the event in a report.

Although automated analysis of text-based reports is faced with great challenges, we believe that by taking advantage of the advanced techniques developed in the field of natural language processing it is a solvable problem. Additionally, the effort is justified in light of the great potential insights and benefits it will bring about. Further research aimed at addressing the aforementioned challenges is being undertaken by the authors.

**Acknowledgements**

This research is being performed using funding received from the DOE Office of Nuclear Energy's Nuclear Energy University Programs. The authors would also like to acknowledge the efforts of the undergraduate students involved in this research: Michael Celesti and Charles Cummings for their assistance in analysing the reports and summarizing the rules; and Edillower Wang for his work at the beginning of the project and the introduction of the Stanford CoreNLP API.

**References**

[1] NRC: Licensee Event Report Search (LERSearch), (n.d.). https://lersearch.inl.gov/Entry.aspx (accessed March 20, 2018).

[2] J.-L. Chang, H. Liao, L. Zeng, Human-System Interface (HSI) Challenges in Nuclear Power Plant Control Rooms, in: Human Interface and the Management of Information. Information and Interaction, Springer, Berlin, Heidelberg, 2009: pp. 729–737. doi:10.1007/978-3-642-02559-4_79.

[3] Y. Zou, Z. Xiao, L. Zhang, E. Zio, J. Liu, H. Jia, A data mining framework within the Chinese NPPs operating experience feedback system for identifying intrinsic correlations among human factors, Annals of Nuclear Energy. 116 (2018) 163–170. doi:10.1016/j.anucene.2018.02.038.

[4] Z. Šimić, B. Zerger, R. Banov, Development and first application of an operating events ranking tool, Nuclear Engineering and Design. 282 (2015) 36–43. doi:10.1016/j.nucengdes.2014.11.035.

[5] J. Park, Y. Kim, W. Jung, Use of a Big Data Mining Technique to Extract Relative Importance of Performance Shaping Factors from Event Investigation Reports, in: Advances in Human Error, Reliability, Resilience, and Performance, Springer, Cham, 2017: pp. 230–238. doi:10.1007/978-3-319-60645-3_23.

[6] J. Pence, Z. Mohaghegh, C. Ostroff, V. Dang, E. Kee, R. Hubenak, M.A. Billings, Quantifying organizational factors in human reliability analysis using the big data-theoretic algorithm, in: International Topical Meeting on Probabilistic Safety Assessment and Analysis, PSA 2015, American Nuclear Society, 2015. https://experts.illinois.edu/en/publications/quantifying-organizational-factors-in-human-reliability-analysis- (accessed March 19, 2018).

[7] X. Zhang, S. Mahadevan, X. Deng, Reliability analysis with linguistic data: An evidential network approach, Reliability Engineering & System Safety. 162 (2017) 111–121. doi:10.1016/j.ress.2017.01.009.

[8] P. Hughes, D. Shipp, M. Figueres-Esteban, C. van Gulijk, From free-text to structured safety management: Introduction of a semi-automated classification method of railway hazard reports to elements on a bow-tie diagram, Safety Science. (n.d.). doi:10.1016/j.ssci.2018.03.011.

[9] A.H. Rao, K. Marais, High risk occurrence chains in helicopter accidents, Reliability Engineering & System Safety. 170 (2018) 83–98. doi:10.1016/j.ress.2017.10.014.

[10] NRC: 10 CFR 50.73 Licensee event report system., (n.d.). https://www.nrc.gov/reading-rm/doc-collections/cfr/part050/part050-0073.html (accessed March 20, 2018).
[11] D. Jurafsky, J.H. Martin, Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, Prentice Hall, Upper Saddle River, N.J., 2000.
[12] M.P. Marcus, M.A. Marcinkiewicz, B. Santorini, Building a large annotated corpus of English: the penn treebank, Computational Linguistics. 19 (1993) 313–330.
[13] Marie-Catherine de Marneffe, Christopher D. Manning, Stanford typed dependencies manual, Stanford University, 2008.
[14] Sebastian Schuster, Christopher D. Manning, Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks - Semantic Scholar, in: 2016. /paper/Enhanced-English-Universal-Dependencies%3A-An-for-Schuster-Manning/7c1a11fcc0d4aa8d7fcfcfa8e375f31f8f23c77a (accessed March 21, 2018).
[15] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural Language Processing (Almost) from Scratch, Journal of Machine Learning Research. 12 (2011) 2493–2537.
[16] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky, The Stanford CoreNLP Natural Language Processing Toolkit, in: Association for Computational Linguistics, 2014: pp. 55–60. doi:10.3115/v1/P14-5010.
[17] D. Bahdanau, K. Cho, Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, ArXiv:1409.0473 [Cs, Stat]. (2014). http://arxiv.org/abs/1409.0473 (accessed March 16, 2018).
[18] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to Sequence Learning with Neural Networks, in: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014: pp. 3104–3112. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf (accessed March 16, 2018).
[19] S. Jean, K. Cho, R. Memisevic, Y. Bengio, On Using Very Large Target Vocabulary for Neural Machine Translation, ArXiv:1412.2007 [Cs]. (2014). http://arxiv.org/abs/1412.2007 (accessed March 16, 2018).