Learning From Text-Based Close Call Data

P. Hughes*, M. Figueres-Esteban &, and C. van Gulijk
University of Huddersfield, Huddersfield, United Kingdom

Abstract: Moving away from standard approaches of safety risk analysis to new approaches that incorporate big data analytics brings with it many opportunities to include new sources of data. These data sources could be the numeric data sources that are used for traditional safety analyses, but could also include text-based sources, such as accident reports, or even social media data feeds. This paper describes an automatic text mining approach to obtain information from close call events (accident "near misses") that can be used for safety management decision-making. The results from this work have shown how automated text mining can be used to extract information from big data sources and be used to inform safety decision-making. Further research in this area intends to look at how the techniques that have been proven to date can be improved with the use of machine-learning techniques.

Keywords: Close Calls, Natural Language Processing, Railway

1 INTRODUCTION

1.1 Close Calls

The GB railways maintain the Close Call database to describe hazardous situations where an event sequence could lead to an accident if it had not been detected, interrupted by a planned intervention or by random event (Gnoni, Andriulo et al. 2013, Andriulo and Gnoni 2014). Workers within the GB railway industry are able to voluntarily report such events on the railway by submitting data to the Close Call database. Close Calls reporting is fundamentally different from accident reporting: since there is no loss, there is no clear definition of what is a Close Call. Also accident reports are in some way more limited since they provide data only on incidents where dangerous situations have occurred resulting in some form of damage to people, property or the environment, whereas close call reports do not require damage to have occurred. Despite these differences, there is nevertheless a link between close call reports and accident reports: dangerous situations described in close call reports provide information on hazards that could lead to or may have led to accidents. As such, analysis of close calls can be expected to provide valuable information relating to the precursors of accidents. The GB railways' Close Call Database contains hundreds of thousands of entries which contain safety information that is useful to the railway.

The GB railways' Close Call database consists of a mixture of structured, categorised data and freeform text. There are currently around than 200,000 records in the database. This large amount of textual data presents a challenge for big data analysis: with such large quantities of information it is not practical to read and understand the relevant safety information without the aid of automated techniques, however the textual nature of the data means that numeric analysis techniques cannot be applied. Table 1 shows the free-text descriptions from four sample records; these records have been modified to change information that may identify the people involved in the incident. In addition to the freeform text, each record is also categorised into one of 26 categories for the event; Table 2 shows the categories used in the database.

Table 1. example freeform text data from four Close Call records

Describe the Close Call event	*What to do about it*
Operatives found working on the east platform with insufficient barriers to protect the public from the work site. Members of the public could have strayed into the work area <!-- RICH TEXT -->	Erected barriers to completely close off the work area i.e. close the gaps in the temporary fence line
Sayers Station Machine/crane AAA no identity badge worn. Operatives/Supervision unclear whom was directing RRV.	Provided with crane controller armlet. Supervisor check for ID. Prior to shift. Badge/armlet provided.
reported by A global operative Lap 9X7 - loc 39 (SWB Z84A499)cable damage seen as cable emerge from the trough route. SSL/BCM to rectify asap	photos taken, reported the SSL SHEQ
Don&#65533;t walk By Draws on BBBB project have not happened for over 12 months. Reported on Previous DWB s	PM to restart asap

Table 2. Event categories used in the database

CONTROL OF CHEMICALS
PROTECTIONARRANGEMENTS
CONFINED SPACES
PUBLIC PROTECTION
ECOLOGY (PLANT & ANIMALS)
RAIL VEHICLES
ELECTRICAL SAFETY
RAILWAY OPERATIONS
EXCAVATION SAFETY
ROAD VEHICLES
FIRE SAFETY
SAFE SYSTEMS OF WORK
LIFTING OPERATIONS
SITE HOUSEKEEPING
MOVING PLANT & MACHINERY
HAZARDOUS SUBSTANCES
NUISANCE (NOISE, LIGHTING)
THIRD PARTY INTERFACE
PERSONAL HEALTH
TOOLS AND EQUIPMENT
PERSONAL PROT. EQUIPMENT
WASTE
POLLUTION (DUST, OILS)
WORK AT HEIGHT
PROCESS/DOCUMENTATION
OTHER

1.2 Research Objectives

This paper describes the technique that has been undertaken in analysing the data in the Close Call database in order to uncover information required to manage safety effectively with the intention of preventing accidents. This work forms one avenue of research in the development of a Big Data Risk Analysis approach for the railways which will integrate data from many different sources including incident databases as wells as sources such live data of passenger movements or weather data, to develop an overall view of safety risk for the railways (Van Gulijk et al. 2015).

1.3 Document structure

Section 2 of this paper describes the theoretical background that has formed the basis of this work; Section 3 provides a description of the analysis method employed in the study; Section 4 provides conclusions and directions for further study.

2 LITERATURE REVIEW

2.1 Close Calls

The importance of analysing and learning from close calls has become increasingly accepted by both industry and the scientific community (Bird and Germain 1966, Jones, Kirchsteiger et al. 1999, Dillon and Tinsley 2008, Bliss, Rice et al. 2014). Across a range of industries including aviation, chemical processing, nuclear power and healthcare, management of close calls has become an important area of safety risk management (Davies, Wright et al. 2000, Macrae 2014). There are different definitions amongst researchers about close calls, near-misses and/or weak signals (Dillon and Tinsley 2008). This work uses a definition by Gnomi and Anndriulo which fits the railway industry: a close call is a hazardous situation where the event sequence could have led to an accident if it had not been detected, interrupted by a planned intervention or by random event (Gnoni, Andriulo et al. 2013, Andriulo and Gnoni 2014).

An early concept proposed by Heinrich, and that continues to be upheld amongst safety researchers, is that there are more close call events than loss-producing accidents, and that action taken to prevent close calls will necessarily result in a reduction in accidents (Heinrich 1931, Bird and Germain 1966, Petersen 1971, Wright and Van der Schaaf 2004, Gnoni and Lettera 2012). Given the larger body of data, close calls therefore potentially provide a richer source of safety-relevant data than accident data alone, and analysis of close call data may be used to assist in the development of safety management system more efficiently than by considering only records of accidents. In fact close call event data can be used to detect and prevent the precursors of accidents (Clarke 1998, Jones, Kirchsteiger et al. 1999) and can therefore assist an organisation in improving their safety management practices even where there has been no history of accidents (Andriulo and Gnoni 2014). In this way, close call data can be used to manage the precursors to events that have a low likelihood of occurring but that have potentially very large consequences. Managing close calls events can be seen as a valuable activity in its own right and even as an accolade of an organisation's safety management capability: Jones, Kirchsteiger et al. 1999 found an inverse relationship between the number of reported near misses and the number of accidents, suggesting that the rate of near-miss could be an indicator of an organisation's safety awareness.

Whilst management of close calls can be useful to an organisation, it comes at a price: the potentially large volume of close call data, compared with the volume of safety data, allows for improved decision-making in safety management (Wright and Van der Schaaf 2004), but brings with it an overhead in terms of the additional effort required to analyse the body of close call data (Gnoni and Lettera 2012).

Despite its potential benefits, operating a close call reporting system is not straightforward. The efficacy of a close call reporting system depends on the quality of the reporting system that provides the source data and the degree to which an organisation's employees provide meaningful data. Clarke notes that if there is a lack of trust between an organisation and its employees, there will be unwillingness for events to be reported (Clarke 1998). In cases where anonymity cannot be guaranteed for employees, a close call reporting system can become derelict (Davies, Wright et al. 2000). Therefore involvement of employees to feel safe in reporting events is a critical component of any safety management system (Clarke 1998, Wright and Van der Schaaf 2004, Gnoni, Andriulo et al. 2013).

A further issue when managing close call events is the interpretation of the event itself. By definition a close call event is one that could have – but didn't – lead to a loss; as such it is possible for close call events to be seen as a success of a safety management in preventing accidents. It is essential that organisations avoid a rose-tinted view of their performance if they are to use close calls to enhance their safety management processes (Dillon and Tinsley 2008).

2.2 Natural Language Processing

Whilst analysis of close call data presents a cost to an organisation, if part of the analysis work can be automated by computer analysis, the cost does not have to be unreasonable. Computer-based close call reporting systems can be implemented in a number of ways, for example by recording coded information, allowing free-text information, or a mixture of both. It has been found that use of free-text data entry allows reporters to provide more details of events and surrounding circumstances than coded information, which in turn allows for the identification of new, or previously unknown, risk factors (Taylor, Lacovara et al. 2014). Computer-based analysis of free-text information requires the use of Natural Language Processing (NLP) techniques which have been an emerging area of study over the past two decades in a number of areas such as road safety and medicine (Allen 1994, Wu and Heydecker 1998, Dale, Moisl et al. 2000, Xu, Stenner et al. 2009). Given the proven successes of NLP techniques in these areas, they have formed the basis of this work.

One of the key problems with NLP is the inherent ambiguity that occurs at the lexical, syntactic and semantic levels of the text (Allen 1994). The written language used to describe close call events is characterised by its use of industry-specific jargon, vernacular terms, abbreviations, and irregular grammatical constructions such as lack of punctuation or missing words (Wu and Heydecker 1998). These challenges often require a pre-processing step to cleanse the source data before any semantic construction can occur. Whilst diverse tools and programming languages have been developed to support the semantic extraction process, each of the tools developed to date has been limited to analysing text within a specific domain (Chang, Kayed et al. 2006, Yafooz, Abidin et al. 2013).

To address domain-specific issues in free-text, a common approach is to develop ontology to represent domain knowledge (Easton and Roberts 2010, Mohan and Arumugan 2011, Verstichel, Ongenae et al. 2011, Hoinaru, Mariano et al. 2013). Using an ontology, key words an terms in the source text are identified and tagged or tokenised to provide short data streams from which information can be extracted (Chandrasekaran, Josephson et al. 1999, Embley, Campbell et al. 1999, Shin, Lee et al. 2014). Such processes invariably result in large quantities of tokens being created which, in turn, can lead to ambiguity in the text. To address this problem clustering methods can be employed to organise text into groups that can be more easily processed (Cui, Liu et al. 2011).

There are a number of off-the-shelf NLP software tools available to support NLP tasks. These tools perform various functions of NLP including: classifying documents, extracting information, providing search engines, and translating text.

Whilst the specific components provide powerful tools for language analysis, in some cases they lack the flexibility to control the workflow of the analysis and can require users to perform data manipulation to ensure that the data are in the correct format to enter into the components or to transport data between them. The generic frameworks, conversely, provide the user with greater control over the framework, but at the cost of requiring more configuration effort. For this work it was decided to use generic components, coupled with computer programming tools to facilitate automation of the analysis.

3 PROCESSES FOR ANALYSIS OF THE CLOSE CALL DATABASE

Bringing together the techniques required for NLP processing of close call data, an overall workflow for extracting information from freeform text in Close Call records was created that involves five processes:

Process 1: Text cleansing, tokenising, and tagging
Process 2: Ontology parsing and coding
Process 3: Clustering
Process 4: Text analysis and method refinement
Process 5: Information extraction

An overview of the process is shown in Figure 1. The techniques developed to carry out these processes are described in Sections 3.1 to 3.4.

To perform this work, 12 months' data from the Close Call database was obtained from the close call website; this data source contains data that has been cleansed to remove personal data. The data describe events that occurred between 01 September 2013 and 31 August 2014; in total there were 72,607 records.

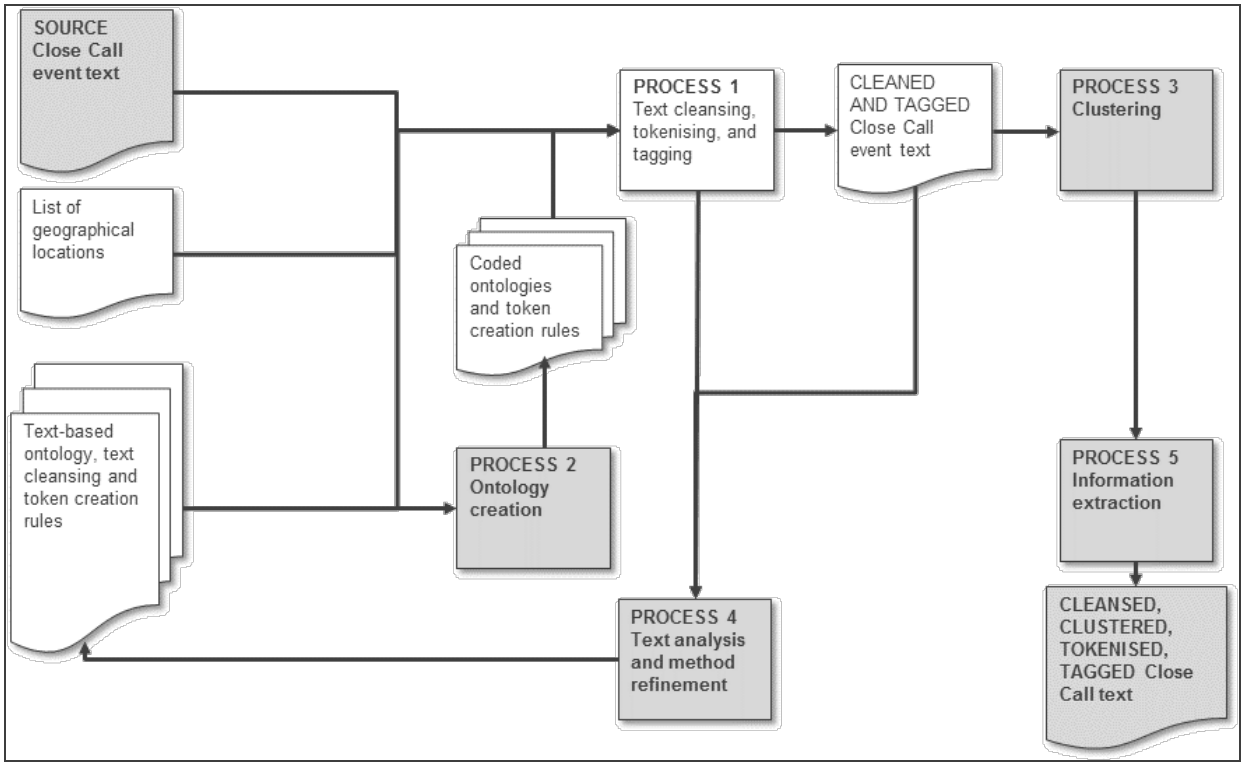

Figure 1: Overview of process of analysis of the close call database

3.1 Process 1: Text cleansing, tokenising, and tagging

3.1.1 Text cleansing

The dataset imported from the Close Call database contains a number of coding artefacts that are not needed for the work of semantic extraction. The example records in Table 1 include the following examples of text: `<br /> <div> </div> <!-- RICH TEXT -->`, and an occurrence of the character ◆ in place of an apostrophe ('). These data artefacts are unnecessary since they do not provide information useful for safety analysis and can interfere with NLP processes. As a pre-processing step software was developed to clean up the database: the software imports the source data, searches for occurrences of specific unwanted text elements and either removes the unwanted text or, if it is possible to determine with a high degree of reliability what the correct text should be, substitutes the text with corrected data. The cleansed text is exported to a new file.

3.1.2 Tokenising

A token is a word or symbol that is used in place of commonly used words or terms. Since the Close Call database has been created specifically to report safety-related incidents on the railway, there are a number of words and terms that specifically relate to railway safety. To ease analysis of the text, these words and terms are identified and replaced with tokens that can be readily identified and used within the text processing.

For example, the term British Transport Police occurs commonly in the source dataset. During the data processing this commonly occurring three-word term was identified and replaced with the single token BRITISH–TRANPORT–POLICE–. Similarly other commonly used terms that mean the same thing can be identified and replaced with the same token, for example instead of British Transport Police, it is common that Close Call authors will write: BTP, or B.T.P., or sometimes BTPolice. These occurrences were all replaced with the same token allowing all instances of these words that all mean the same thing to be grouped together as necessary.

3.1.3 Spelling checking and correcting

Since the data considered during this study is freeform text, there are many variations of spellings used in the source data. Table 3 shows examples of spelling variations identified for the word palisade.

Since the process of tokenising requires identification of words for conversion to tokens, spelling variations reduce the efficiency of this process. As part of the process of text cleansing, two automatic spelling checking and correcting tools were used to identify and, where possible, automatically correct words that were entered with non-standard spelling. The spelling checkers used during the study were the spelling checker that is included in the Python regular expressions language extension, and the spelling checker that is included as part of the Microsoft Office range of software.

Table 3. Identified alternative spellings of the word palisade

palasaid	paliside
palasaide	palistrade
palicade	pallasade
palilsade	pallasde
palisadade	pallaside

3.1.4 Tagging

A large amount of the text in the Close Call database is useful for determining the details of an event such as place names (for example: Winchester, Newcastle), distances (for example: 24 miles 6 chains), and times (for example: 03:19 hours). Whilst these data are useful for understanding the details of an event, they can complicate the automated process of text analysis. To address this issue common sequences of text that relate to places, distance and times were identified. To identify places, a lists of British railway stations and names of British towns were obtained. Software was developed to automatically search for words within the Close Call database that matched names in these lists. Once a place name had been identified, it was extracted from the record and replaced with the token [PLACETAG]. The place name was then added as supplementary data (a tag) in another field attached to the original record. In this preliminary study, to avoid unintentionally changing the meaning of a record, place names that are also English words (such as Bath, Hove, Reading, Stoke, Beer) were removed from the search list and therefore were not removed from the record.

Distances along the railway are often recorded in the Close Call database by use of identifiable patterns of text such as (number) miles (number) yards and the various abbreviations that are used for units of distance such as: m for miles, ch for chains, or yds for yards. Similarly, times were identified using a similar process to identify commonly occurring patterns of text such as 14:29, 2.29 pm, at 1429 hours. The tagging process identified appropriate phrases and replaced them with a token: distances were replaced with [DISTTAG], and times were replaced with [TIMETAG]. As with identified place names, identified distance and time data were appended in a separate tag field.

3.2 Process 2: Ontology parsing and coding

An ontology is a group of concepts that are related in some way; these concepts are usually expressed in words or terms. An ontology for railway descriptors of hazards would define the hazards and controls that are associated with the railway and the relationships between the hazards. For example,

within railway operations there are numerous hazards that could affect members of the public at a station such as platforms, walkways, turnstiles, stairways, steps handrails, and escalators.

An ontology lists these elements in a way that shows the relationships between them. During the search of published material, no existing ontology was identified that would be suitable for assessing data in the Close Call database. Therefore an ontology was developed by considering each of the identified tokens and, where possible, describing conceptual associations between them. An ontology for railway hazards provides a tool for clustering records that have similar meaning even if different words or terms have been used to described the hazards.

3.3 Process 3: Clustering records

An ontology provides a tool that makes it possible to cluster records in the Close Call database based on the proximity of the relationship of their tokens in the ontologies. To perform clustering of records, this work proposes a process that considered all the tokens within a pair of records, and then uses the ontology to determine the proximity of the tokens in one record to the proximity of the tokens in another record. To provide a measure of the proximity of two tokens in the ontology, this work also proposes a scoring system that is based on the genetic proximity between members of a family, as shown in Figure 2.

For the initial test of the clustering process, the proposed scoring system shown in Table 4 was used to rate the proximity of relationships between tokens.

Figure 2: examples of family relationships relative to the First Node

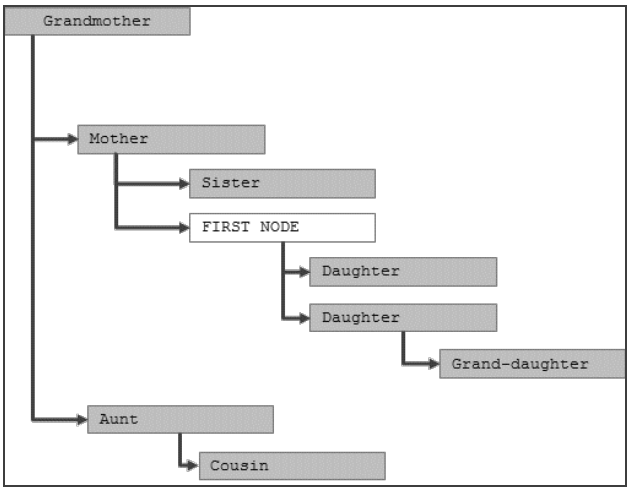

Table 4: proposed relationship proximity scores

From first node to	*Strength score of relationship*
identical token (twin sister)	1.000
Sister	0.500
mother	0.500
grandmother	0.250
daughter	0.500
grand-daughter	0.250
Aunt	0.250
Cousin	0.125

To progress from considering the proximity of two tokens, to considering the overall proximity of two records, the proximity scores of each pair of tokens in both records was summed. The resulting score gives an indication of the semantic similarity of a pair of records, and by considering the similarities of a group of the entire dataset, records that are similar in meaning, but that use different words, can be clustered into groups of semantically similar records.

However, without normalising the proximity scores two records will be higher when they simply have a larger number of tokens: a record that contains a large number of tokens will have a higher chance of obtaining a high score with any other record compared to a record that contains only a small number of tokens. As such it can be expected that records that are more verbose will contain more tokens and therefore will obtain higher proximity scores with any other record. These higher proximity scores obtained in this way will not be a function of the semantic similarity of the records, but rather a function of the verbosity of the records; it is therefore necessary to provide a method of normalisation so that the proximity score is more representative of the actual semantic similarity between two records. This work proposes a normalisation method that uses a calculation of the proximity score that

a record obtains between itself and an identically worded record (in effect the proximity relationship score with itself) and the proximity score it obtains compared with another record. The proposed proximity scoring method is:

$$P_{(m,n)}=(2p_{(m,n)})/(p_{(n,n)}+p_{(m,m)})$$

Where:

m and n identify two individual Close Call records that are compared;
$P_{m,n}$ is the normalised total proximity score between records m and n;
$p_{n,n}$ is the un-normalised proximity score between record n and itself;
$p_{m,m}$ is the un-normalised proximity score between records m and itself.

The proximity scores are used to identify records that are semantically similar and to create clusters of records that are all similar in meaning and are distinct from other records. These clusters can then be examined to identify trends such as time or location of occurrence.

3.4 Process 4: Text analysis and method refinement

Once the process of clustering has been completed, it is possible to use automated searches of words within each cluster to identify words that occur commonly and occur commonly together (colocations). The colocation review employed Pointwise Mutual Information measure (Church et al., 1990) to identify commonly occurring words as well as pairs and triplets of words that commonly occur next to each other.

A manual review of the colocation results was performed to identify additional words, bigrams or trigrams that need to be included as tokens, and to identify other text cleansing rules (Process 1).

3.5 Process 5: Information extraction

Processes 1 to 4 provide a method to automatically process the text-based data in the Close Call records. These processes are relatively straight-forward and can automatically be applied to all records in the database with minimal or no manual intervention. The final process, information extraction, is less straight-forward since it attempts to provide answers to questions that are relevant to management of safety on the railways. This work proposes that automated information extraction would require a process that can identify meaningful events or trends in Close Call text, either by automatic pattern-detection processes or in response to questions formulated by an analyst; and present results to an analyst in a way that supports the decision-making process for improving safety on the railways.

This process of automatic information extraction was not codified as part of the work described in this report. Instead this study has used two simple case studies to demonstrate how the process proposed in Sections 3.1 to 3.3 can be used to provide learning from Close Call records. Further work would involve the development of automated processes that can perform the tasks undertaken in the case studies. The questions considered in the case studies were:

Question 1: The SMIS database shows that near miss incidents with track workers take place more frequently between the hours 10:00 and 14:00; is this pattern also present in the Close Call database?

Question 2: Do trespasses take place at certain times of the day or do they take place with equal frequency throughout a 24-hour period?

4 RESULTS

The questions were analysed manually: data was extracted from the Close Call database using the techniques described in Sections 3.1 to 3.3 and analysed.

To identify Close Call records relating to track worker near miss incidents a NEAR–MISS– token can be created whenever the following terms are found in a Close Call record: `near miss`, `nearly hit`, `nearly struck`. Records containing this token were identified and analysed to determine the time of day at which events occurred. These results were compared with the data in the SMIS data to determine if there is a correlation between the times of day at which these events occur. As a further test, a comparison was made between the near miss events in the accident database and all events in the Close Call database. The correlation between the distributions was tested using a chi-squared analysis.

To address Question 2 (at what time of day do trespass events occur?) records in the Close Call database can be identified by creating a TRESPASS– token whenever a record was found to contain specific text such as: `person on track`, `people on track`, `trespasser`, `trespassing`. A search was then performed to identify Close Call records that contained the TRESPASS– token and extract time tags where these are available. Overall it was found that similar trend exists between the Close Call and SMIS data, however there are some significant differences between the two datasets that require further investigation.

5 CONCLUSIONS

This work has proven the effectiveness of the procedures that would be required to perform rudimentary analysis of the text; for example to identify types of events, such as the times of day when track worker near miss close calls occur and when trespass events occur. In this report, the method for extracting safety lessons is relatively crude: pre-determined safety questions have been investigated by analysing the Close Call database with the NLTK processes. This work demonstrates an approach that can be used to extract safety lessons from a text-based big data source: the Close Call database. Further work is required to develop more complex search processes to unlock more subtle safety lessons from the Close Call database, however there is no reason to assume that this is impossible.

The work described in this report describes the basis for one of a suite of tools that will be required to obtain safety-relevant information that can support decision-making for the railways. The creation of an effective process for text mining safety information from big data is one step in the overall process of integrating different sources of information (for example text-based incident data could be combined with numeric data on train movements or passenger activity, or even video-based data) to provide an overall picture of safety risk on the railways. Bringing together different sources of safety data in this way forms the basis of Big Data Risk Analysis with the aim of identifying safety risks on the railway that could either not previously be identified, or could not be identified in a timely manner.

References

Allen, J. F. (1994). Natural language processing. Encyclopedia of Computer Science, John Wiley and Sons Ltd.: 1218-1222.

Andriulo, S. and M. G. Gnoni (2014). "Measuring the effectiveness of a near-miss management system: An application in an automotive firm supplier." Reliability Engineering & System Safety 132(0): 154-162.

Bird, F. E. and G. L. Germain (1966). Damage control: A new horizon in accident prevention and cost improvement, American Management Association New York.

Bliss, J. P., S. Rice, G. Hunt, K. Geels (2014). "What are close calls? A proposed taxonomy to inform risk communication research." Safety science 61: 21-28.

Chandrasekaran, B., J. R. Josephson, V. R. Benjamins (1999). "What are ontologies, and why do we need them?" IEEE Intelligent systems 14(1): 20-26.

Chang, C. H., Kayed, M., Girgis, M. R., & Shaalan, K. F. (2006). “A survey of web information extraction systems. Knowledge and Data Engineering”, 18(10), 1411-1428.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational linguistics, 16(1), 22-29.

Cui, J., Liu, H., He, J., Li, P., Du, X., & Wang, P. (2011). Tagclus: a random walk-based method for tag clustering. Knowledge and information systems,27(2), 193-225.

Dale, R., Moisl, H., & Somers, H. (Eds.). (2000). “Handbook of natural language processing”. CRC Press.

Davies, J. B., Wright, L., Courtney, E., & Reid, H. (2000). Confidential incident reporting on the UK railways: The ‘CIRAS’System. Cognition, Technology & Work, 2(3), 117-125.

Dillon, R. L. and C. H. Tinsley (2008). "How Near-Misses Influence Decision Making under Risk: A Missed Opportunity for Learning." Management Science 54(8): 1425-1440.

Easton, J. M. and C. Roberts (2010). Railway Modelling: The case for ontologies in the rail industry. In Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD). Valencia, Spain: pp. 257–262.

Embley, D. W., Campbell, D. M., Jiang, Y. S., Liddle, S. W., Lonsdale, D. W., Ng, Y. K., & Smith, R. D. (1999). Conceptual-model-based data extraction from multiple-record Web pages. Data & Knowledge Engineering, 31(3), 227-251.

Gnoni, M. G. and G. Lettera (2012). "Near-miss management systems: A methodological comparison." Journal of Loss Prevention in the Process Industries 25(3): 609-616.

Gnoni, M. G., Andriulo, S., Maggio, G., & Nardone, P. (2013). “Lean occupational” safety: An application for a Near-miss Management System design. Safety science, 53, 96-104.

Heinrich, H. W., (1931). Industrial accident prevention, McGraw-Hill New York.

Hoinaru, O., Mariano, G., & Gransart, C. (2013, January). Ontology for complex railway systems application to ERTMS/ETCS system. In FM-RAIL-BOK Workshop in SEFM’2013 11th International Conference on Software Engineering and Formal Methods.

Jones, S., Kirchsteiger, C., & Bjerke, W. (1999). The importance of near miss reporting to further improve safety performance. Journal of Loss Prevention in the process industries, 12(1), 59-67.

Macrae, C. (2014). Close Calls: Managing Risk and Resilience in Airline Flight Safety, Palgrave Macmillan.

Mohan, A. and G. Arumugan (2011). "Constructing Railway Ontology using Web Ontology Language and Semantic Web Rule Language." Int. J. Comput. Tech. Appl 2(2): 314-321.

Petersen, D. (1971). Techniques of safety management, McGraw-Hill New York.

Shin, Y., Lee, S. J., & Park, J. (2014). Composition pattern oriented tag extraction from short documents using a structural learning method. Knowledge and information systems, 38(2), 447-468.

Taylor, J. A., Lacovara, A. V., Smith, G. S., Pandian, R., & Lehto, M. (2014). Near-miss narratives from the fire service: A Bayesian analysis. Accident Analysis & Prevention, 62, 119-129.

Van Gulijk, C. et al., 2015. Big Data Risk Analysis for Rail Safety? In Proceedings of the 25th European Safety and Reliability Conference, ESREL 2015.

Verstichel, S., Ongenae, F., Loeve, L., Vermeulen, F., Dings, P., Dhoedt, B., Turck, F. D. (2011). Efficient data integration in the railway domain through an ontology-based methodology. Transportation Research Part C: Emerging Technologies, 19(4), 617-643.

Wright, L. and T. Van der Schaaf (2004). "Accident versus near miss causation: a critical review of the literature, an empirical test in the UK railway domain, and their implications for other sectors." Journal of Hazardous Materials 111(1): 105-110.

Wu, J. and B. G. Heydecker (1998). "Natural language understanding in road accident data analysis." Advances in Engineering Software 29(7–9): 599-610.

Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R., & Denny, J. C. (2010). MedEx: a medication information extraction system for clinical narratives. Journal of the American Medical Informatics Association, 17(1), 19-24.

Yafooz, W., Abidin, S. Z., Omar, N., & Idrus, Z. (2013, November). Future trends in managing extracted information. In Control System, Computing and Engineering (ICCSCE), 2013 IEEE International Conference on (pp. 279-283). IEEE.