MACHINE LEARNING APPLIED TO DATABASES TO IMPROVE SAFETY

E.Cox. Author lead presenter[1], G. Foeillet. second presenter[2], and J. Ferreira, third presenter[3]

[1] *Head of SNCF RESEAU /DG_NUM_DAD,* emmanuel.cox@reseau.sncf.fr

[2] *SNCF /DRASS* In charge of Safety affairs, guillaume.foeillet@sncf.fr

+33 6 26 22 73 97

[3] *SNCF RESEAU /DG_NUM_DAD_S4.0,* jeremie.ferreira@reseau.sncf.fr

BACKGROUND

For years, Railway companies have built safety event reports, most of the time stored in flat files in some cases in databases. Event reports may have pre-defined (*structured*) or free-text (*unstructured*) formats; can even also be found scans of hand-written pages, and also anything related to event useful for safety enquiry, such as pictures from the safety event scene, drawings explaining how things occurred chronologically, newspaper articles witness the progress of investigation and setting of the conclusions. Voice recording and its text conversion from both ground railway traffic controller and train driver are even considered for the future.

Storing these data corresponds usually at the beginning to a National commitment for data traceability, requiring to keep in a safe place all types of documents relative to safety event data. Since then, this is used from time to time to retrieve all details from one specific safety event, either because there is some progress in the investigation, or because it appears similar to a newly occurred safety event, and there may be some benefit in having a look at it. But these archives are seldom used comprehensively to perform efficient data mining.

The reason for that is that attempts to obtain some safety benefits from such parsing though archives are rarely fully conclusive. This is mainly due to the fact that storage tools have not been initially designed to perform such efficient research, they do not provide the means to find out all the useful data and also because those who generate the reports shortly after the safety events occurred are not those who could later-on have analyse them to make conclusions. Data are indexed by date, sometimes classified according to safety event taxonomy. Parsing tools may propose an embryo of data filtering, based on exact field string matching, which cannot be productive since one does not always know what were the exact words chosen for the safety reports. There could be some results, when by chance provided words match exactly, but the users are never sure that the complete set of reports has been extracted, precisely because these provided words may not be always the same used in all the records that could be interesting to retrieve for analysis.

Two examples of partially successful / hardly conclusive requests are exposed hereafter. The poor results, the amount of time spent to check manually what the computer was unable to deliver automatically have led to the decision to launch de design of this new kind of tool.

1rst topic : Analysis after an accident at crossing level, leading to loss of train radio communication

At the vicinity of a crossing level, a regional train is traveling at 136km/h. Driver suddenly sees that a farm tractor seems to have his wheels stuck in track at the crossing level. Whistle is blown and emergency braking is activated, but the driver does not have the time / the idea to send a radio alert The farm tractor driver decides to get off after whistle is blown again. Collision cannot be avoided; train hits the tractor at a speed of 118km/h. Tractor is detached from its load and falls aside. The two front bogies of the trainset have derailed. The train driver is injured but conscious and sees that the train has stooped within the gauge of the adjacent track. To avoid an additional collision with another train, driver tries to send a radio alert, but it does not work. It will later be established that the electrical wiring has been damaged close to the crush of the automatic couple during the collision. This has led to the idea that a thorough design review on all types of rolling stock, in order to determine if there was some weakness in the electrical design, therefore to be improved, and to determine if some part of fleet was weaker on that particular aspect, that would need some top priority modification.

It has been suggested to parse the database of safety events to determine all similar event (collision at crossing level) in order to collect what type of train had been involved, and determine if that could be useful to obtain some answers.

➔ Events at crossing level have been quite easy to collect, but it has appeared that due to free text description, and lack of precise request of filling-in such data in the structured file, it has been impossible for the computer to retrieve the exact type of rolling stock involved. More than 200 files remained unclassified; it has been necessary to open manually each of them, to try to find out some valuable information whenever it could be found. it remained unconclusive for some of the proposed files, for instance du to the fact that the description of the safety event only refers to "the train", not providing precise information about the rolling stock type. The safety event itself is at the center of the concern, and no one could foresee at that time that the type of rolling stock would be an important information to mention. Once the filtering has been finalized 'by hand', the analysis has been made, and the conclusion was that there was no model of rolling stock that appears to have a weaker design than the others. It is only because there was the employee's will to spend enough time to finish the filtering by hand that the analysis did take place after that; the computer was not helpful enough, could not take care of such tedious job.

2nd topic: [Cost / efficiency] ratio for a proposed investment, to significantly reduce the number of SPADS

An equipment manufacturer came to train operator company to present a prototype of hardware + software system using cameras and other sensors to detect and recognize all trackside signposts and signaling lamps. the purpose of the product is to provide to train drivers an information device in order to reduce drastically the cases where drivers miss the signposts and pass signals at danger (SPADs). This appeared to be a very interesting perspective to provide such a new equipment that would help the driver whenever it would be useful to do it, and only when useful, that is when obviously the driver does not show any sign of taking the trackside information in consideration (reducing speed, or preparing to stop at appropriate distance.

It has been suggested to parse the database of safety events to determine all similar events (all the safety events due to SPADs) in order to collect whether casualties, wounds, damages (consequences) could have been avoided if such a system had been available, and determine what is the [cost / efficiency] ratio of the proposed product in order to decide or not to launch its design.

➔It has been easy to collect the safety events where SPADs could be found, but it has appeared that due to free text description, it has not always been possible collect precisely what were the consequences of each of them. It can partially be explained by the fact that when safety event documents are written, immediately after it has occurred, not all the consequences are known, the person in charge of the report clearly does not have any idea of how much repair of the damages could cost, furthermore when the investigation is finished and the experts have provided their financial estimates, it is not added into the safety event description. This makes sense since it is not a safety topic but a financial topic, handled after the event by other specialists.

Many of the extracted events could be considered as useful to establish the [cost / efficiency] ratio, and those which have been used contained vague information about costs of consequences (in most cases, it is asked to say whether of not it less than of more than threshold financial values, never updated with exact values later on, this information remaining in the hands of other teams than the safety employees.

Although very frustrating, it has been possible to conclude with enough approximate evidence, that the [cost/efficiency] ratio for the product was quite low, leading to the decision that the safety Department on itself could not decide to launch such a product, because it concerned too few instances among safety events and would not be clearly considered as a top priority efficient product, as compared to other investments to improve the railway safety.

(Many other examples of attempts to make use of the safety event databases could be considered…)

OBJECTIVE

It is obvious that the safety event databases do contain interesting data, and to make these data truly more valuable, SNCF's has decided to combine semantical analysis [i.e., determining the exact meaning of the provided set of words, in order to be able with it afterwards to find matches with the same or approaching meaning of the words which are found, which can be different ones], with Natural Language Processing (NLP) [using computing techniques to analyse efficiently text and determine semantical similarities during an automatic comprehensive parsing].

The resulting 'Machine learning software' can easily be linked to any data storage; once key words provided, the search engine parses all information and acts as if information quality was improved, without truly modifying it: unstructured data becomes hence as useful as if it was structured. Safety reports could also be generated in semi-automatic mode.

It has been initially designed for railway safety purposes, but it could be used for any railway domain other than safety, and could even be used also by other sectors that railway.

METHODS

Preparing the search engine

Core functions of the search engine. They do not vary when a query is entered.

Those functions have been tested on a corpus of data consisting of approximately 800 archived incident reports (namely RACs). The data corpus has been cleaned and pre-processed according to different NLP methods, such as "tokenization" (a process by which a piece of sensitive data, is replaced by a surrogate value, the latter known as a "token"). The stopwords (words that are too common to be of any semantic significance and efficient discrimination use) have been deliberately removed and are therefore not taken (never) in account by the search engine.

- **For the construction of the Thesaurus:** (The thesaurus is a lexical network)
    - We have put together the words from the same lexical field based on their 'semantic similarity score'
    - We have calculated semantic similarity between the words in our corpus using KeyedVectors (from the Gensim library)
    - KeyedVectors has been assigned a pre-trained French word embeddings model, frWac2vec
    - We have built word embeddings (put them together); these are vectorial representations, that assign a similar representation to words of similar meaning
    - We have then enriched this first version thesaurus with all possible vocabulary content that is specific to railway domain. This vocabulary

content has been established through documentation often used as reference for such purpose and also the inputs of different railway domain experts, therefore representing the best possible reality of the context, reflecting its technical complexity

- **Fuzzy Search:** Fuzzy Search is a widely used method, based on "Cosine similarity"
- We have used for that the specific library FuzzyWuzzy
- It enables the user to find approximative matches for a specific query, rather than only those matching the exact pattern provided
- also, a threshold of similarity can be assigned for the two patterns to be considered as close matches
- Therefore, this can also help to bypass any issues due to spelling mistakes, or minor variations of a specific word (singular/plural, or feminine/masculine)

Preparing the query :Once a query is entered:

It is more efficient to convert it into a 'regular Expression'

- The query is reformulated in the form of a regular expression
- Regular expressions are a sequence of characters specifying an optimal search pattern
- They are built on the basis of regular grammar
  ➔The first reason for which we have decided to use regular expressions, is that they can take into consideration the distance between the different words of a query, within the provided query, since they appear as being separated by an amount of characters or tokens. For instance, the query constitutes of the words X Y; the expression (X.{ 1,10} Y) indicates that X and Y can be separated by anywhere from 1 to 10 characters
  ➔The second reason for which we have decided to use regular expressions, is that for every word that constitutes the query, we can easily retrieve its (exact or approaching matches, stemming from the thesaurus or from the 'fuzzy' search. For instance, a word X1, has a semantic match X2 and a fuzzy match X3. The expression (X1|X2|X3) enables us to search for X1 OR X2 OR X3 simultaneously

Preparing/Organizing the results

After the query converted into a regular expression has returned results from the corpus, it is necessary to perform:

- **Calculation of semantic similarity**
  - Every resulting match is compared with the original query
  - A semantic similarity score is computed, using the original query as reference
  - The matches are then reorganized according to their semantic score, obtained from calculation
  - Semantic similarity calculation is performed based on the previously presented thesaurus

- **Filtering**
  - Any kind of useful filter can be applied on the results from the corpus, they are based on metadata; e.g., it could be date of event or a threshold one or any designation of a regional location
  - Each document in the corpus contains metadata that differs from its textual content, and is stored in the format of a data frame, making it easy to retrieve
  - We have used the Pandas library in order to handle efficiently those data frames

RESULTS

Results of extraction after all post processings can be put together, the packet being converted as a zip file to paste in a convenient format for the user all the resulting data generated by the tool. This enables the user to perform additional post-processing analysis, by hand or with any other tools, what is for sure is that these data are relevant with the received query.

Potential applications foreseen for safety with this product are huge. Connection to other databases, for instance railway description data, could be made with same approach to demonstrate versatility of product and similar benefits, provided that a customised Thesaurus is first built for each corresponding application.

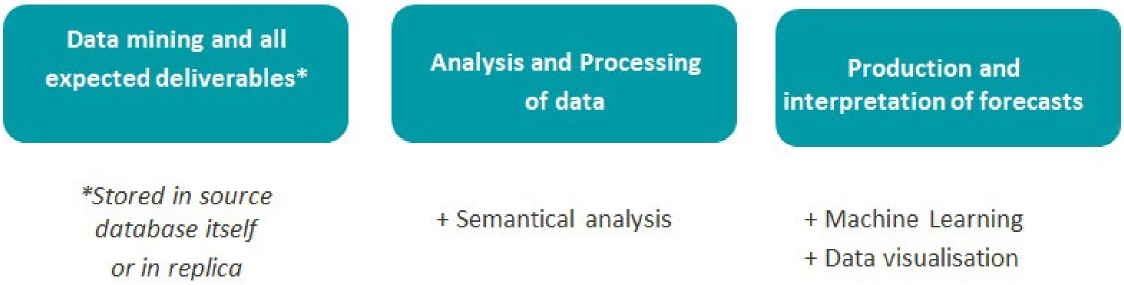
CONCLUSION

(At this stage of our proof of concept, the design and tests will continue on a large scale of data, to confirm and improve the features of the product)

Such a tool can be considered as what is commonly designated by the commonly accepted name 'browser', although it is different from well-known web browsers, since it requires some preliminary work to adjust the tool to the type of domain and the type of queries, so it could not be considered as a unique product that fits all needs; it is efficient once it has been customized. A non customized product could work but would clearly be less efficient.

It provides less expensive, less time consuming, more efficient analysis of the archives, with drastically improved completion of the parsing. Of course, it encourages to perform more of such investigations of available data, apply queries and obtain answers with high confidence of utility for decision making, given objective quantitative results obtained, far better than subjective point of view about potential efficiency, which remains to be proven once put in service on the field. It delivers results with exact or close/related matching, crossing criteria, helps performing more accurate statistics over complete data set.

It contributes building predictions, helps managers to improve safety by making appropriate decisions for the most useful and top priority investments, for instance setting risk control measures to suppress safety events of mitigate its consequences, design new expensive safety equipment with a higher confidence that it will be as efficient as initially promoted.



Data mining and all expected deliverables*

Analysis and Processing of data

Production and interpretation of forecasts

*Stored in source database itself or in replica

+ Semantical analysis

+ Machine Learning
+ Data visualisation

Hints of possible improvements in the future

After consolidation of the on-going developments, taking advantages of higher levels of Artificial Intelligence could be considered and included, for instance, it could be convenient to let the initial thesaurus enrich its contents by itself, as well as the amount of vocabulary that could be considered, so that the product could increase its overall performances without any redesign for that.

Today's version of the product aims to demonstrate to decisional safety managers that the more useful and exact predictions and conclusions are expected, the more it is necessary to have access to data of high quality. the product is therefore an incentive to improve the structure of the safety reports and the quality of the stored data. A semi-automatic fill-in form could be proposed as the unique way to input data relative to a safety event, in order to increase the amount of structured data versus the unstructured data. Such a semi-automatic fill-in form would rather represent a separate tool from the proposed browser, that would by the way directly contribute to the performance of the browser, when applied to data of higher quality because better filled in, with the semi-automatic fill-in form. It could also be decided at one moment in time that free text is no more an acceptable way to provide information relative to a safety event; in that case, the unstructured data would be completely suppressed for the new records of safety events. The performance of browser would hence increase, but some unstructured data would remain, since it is not planned to make changes on former archives, they have to remain the way they are

Later on, (after the proof-of-concept phase, it could be decided to improve the metric features; for instance, analysis of the number of occurrences of the words being looked for, analysis of the type of words being looked for, and also how deliberate co-occurrences in the provided words can be detected, and what meaning it can have to help understand better what is being looked for.

Following this presentation, SNCF is ready to discuss with other railway companies willing to work on similar issues in order to reach thanks to this product some equivalent results when connecting to other safety databases, and also progress the matter, if new ideas could rise from the share of what the priority queries would be.

Keywords: [safety reports]; [machine learning]; [data mining]; [statistics]; [forecasts].