Natural Language Processing (NLP) tools for the analysis of incident and accident reports

Christophe Pimm[1], Céline Raynal[1], Nikola Tulechki[1, 3], Eric Hermann[1], Grégory Caudy[2], Ludovic Tanguy[3]

[1]CFH – Safety Data
4 impasse Montcabrier
31000 Toulouse, France
+33 6 85 01 14 87
{hermann;pimm;raynal;
tulechki}@conseil-fh.fr

[2]Air France
45 rue de Paris
95747 Roissy, France
grcaudy@airfrance.fr

[3]Université Toulouse II – Le Mirail / CLLE-ERSS
5 allées Antonio Machado
31058 Toulouse, France
{tanguy;tulechki}@univ-tlse2.fr

ABSTRACT

The human factor field is expected to evolve due to the development of Natural Language Processing tools which allow for new approaches to handle natural language data. In the current project, we use NLP methods to facilitate experience feedback in the field of civil aviation safety. In this paper, we present how NLP methods based on the extraction of textual information from the Air France ASR can contribute to (i) the improvement of the reliability of the coding, facilitating the coding itself, (ii) the analysis of reports regardless of the categorization in order to expand the analysis perimeter and to avoid the inherent limitations of the codification.

Keywords

Natural Language Processing, Analysis of accidents/incidents, Categorization, Textual similarity

INTRODUCTION

Learning valuable lessons from past incidents and accidents has become paramount to the effort to increase safety in any risk-prone activity. Because of this, national or international regulators such as companies like Air France store a large collection of reports for analysis. Manual analysis of these reports is complex and requires lots of resources. Among many information given to describe a security event, there is a description of the facts written in natural language and a codification: values from a predefined taxonomy. Complexity comes from both the task of categorizing the reports (given the number of values, the users' knowledge, etc.), and the task of analyzing the reports from a global point of view (which is a real issue for knowledge management in companies).

Our goal is to develop tools to help the users in these two tasks of coding and analyzing the reports. Thanks to Natural Language Processing methods, a linguistic analysis of the narrative part is done by a computer and it offers a means of access to a collection of feedback data. In this paper, after giving an overview of the reporting system at Air France, we focus on the processing of the textual information allowed by the NLP methods and its applications. We present two applications: (i) how these techniques can be helpful to code reports, i.e. to pick the correct value among a predefined set; (ii) how, given a database of reports, they can be used to identify similar incidents.

INCIDENT AND ACCIDENT REPORTING AT AIR FRANCE

In order to manage large collections of data, it is common practice to categorize individual reports within a certain categorization schema, consisting of a closed set of category values established upon a particular underlying accident model. Examples of such schemas are ICAO's ADREP taxonomy [1] used mostly by national and international regulators and variations of the Bow-Tie Model used mainly within operators' SMS (Safety Management System). For its ASR and CSR[1] analysis, Air France's switch towards an integrated SMS [2] also involves the implementation of a Bow-Tie Model-based schema for incident report categorization.

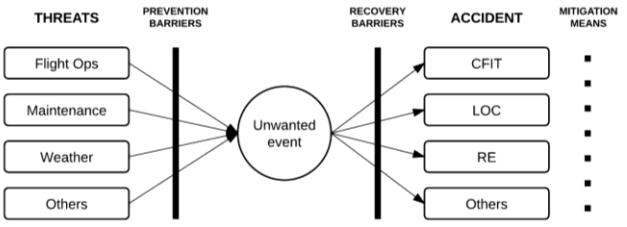

Figure 1: Schematic view of the Bow-Tie Accident Model[2]

The Bow-Tie Accident Model represents a synthetic view of an accident scenario, combining both causal and consequential information. It is centered on the concept of hazard, or "unwanted event" (e.g. "Level Bust" or "Communication Loss"). Once the hazard is identified, a fault tree is built on the left hand side, representing the

[1] Aviation Safety Report, Cabin Safety Report

[2] CFIT: Control Flight Into Terrain; LOC: Loss of Control; RE: Runway Excursion

cause of the hazard in the form of a set of threats which have contributed to it and a set of barriers which have (or have not) prevented these threats from contributing to the hazard (for example "MTO: Turbulence" or "A/C: Noisy Cockpit"). On the right hand side, an event tree is built representing the barriers that allowed recovery from the hazard, as well as the potential accident and the potential mitigation measures that may or may not have been put in place. Hazards which have not occurred due to proper functioning of prevention barriers are also represented.

Categorizing incident reports within this schema requires the coder to choose an item from sets of categories which list all identified threats, barriers, unwanted events, mitigation means and potential accidents (like "CFIT" or "Loss of Control"). Once categorized, individual reports are exploited both in a quantitative way by producing statistics and trends and in a qualitative way, where the categorization is queried to identify and extract individual reports of interest for further investigations.

To facilitate the processing of reports, we suggest using linguistic analysis and Natural Language Processing methods.

AUTOMATIC LINGUISTIC ANALYSIS

By definition, a text written in natural language contains variations: a same idea can be expressed in different ways. To deal with this, a linguistic analysis provides certain standardization and allows for a global processing. That is why the linguistic analysis is an essential prerequisite for subsequent processing. This analysis is based (i) on basic language-dependent processing which can for example be applied in other fields and (ii) on domain-specific processing, a change of domain requiring an adaptation of these processes.

Basic linguistic analysis

The basic linguistic analysis (domain-independent) consists of several processing phases which produce a final list of terms found in the processed narratives and information on these terms (number of occurrences, dependency between them, etc.). It is primordial to emphasize that if a term can be a simple word, it can also be a structured group of words (a syntagm) which will be more relevant for the ensuing analysis. It is more useful to know that a narrative contains the compound term "landing gear" and to extract such a term, rather than consider each of the words separately ("landing" and "gear").

In order to obtain correctly structured compound terms (e.g. "main landing gear" and not "main landing" in such a sentence as "the main landing gear of the aircraft has two wheels"), it is first necessary to assign a grammatical category (verb, noun, adjective, etc.) to each word of each sentence. This process will allow to automatically identify the links between all the words in a sentence and their nature. For a sentence such as "The cabin crew reports a problem" (Figure 2), we can identify that "cabin" qualifies "crew" and that "a problem" is the object of the verb "reports", etc. This dependency between the words in the sentences thus allows to create compound terms (syntagms) which will be automatically rebuilt and extracted. The syntagm reconstruction is performed on the lemmatized forms of the words (infinitive, singular) and not on the inflected forms (conjugated, plural) to harmonize the extractions and isolate the words from the particular form in which they are found.

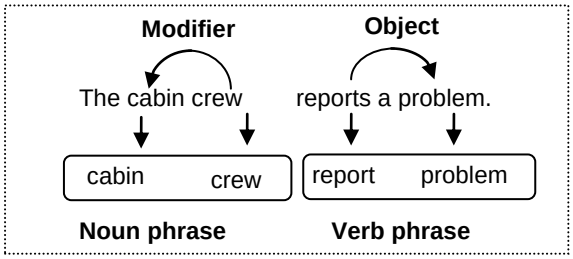

Figure 2: Example of linguistic analysis

It should be noted that some terminological variations are taken into account. For example, phrases which differ in form but not in meaning are grouped under the same term (e.g. "report a problem" and "report the problem" are both grouped under the term "report problem"). The list of extracted terms is saved along with all the occurrences of these terms and the dependency links between them. This information concerning the narrative parts of the reports constitutes the corpus on which the learning process will be performed.

Specific linguistic processing

Processing specific to the aviation domain is added to the basic linguistic analysis we have just described.

- Many acronyms are used in the aviation domain and we want to consider them as equivalents of their developed form. For example, we want all expressions (acronyms or developed forms) referring to automatic pilot ("AP" or "automatic pilot") grouped under the same term, "AP", to prevent semantic information (identical whatever the form used) from being scattered across several terms. To impose this equivalence from the very start of the analysis, a linguistic resource containing all the domain-specific acronyms and corresponding developed forms has been created to facilitate their identification.
- Following the same idea to harmonize the text by grouping forms with similar meaning, another process is performed for quantities. They are detected using a specific resource and standardized for each type of unit: "3 liters" and "15 liters" are recognized and extracted as two occurrences of the same term, "XXX liters". It should be noted that for certain types of units, feet for example, it is relevant to keep some level of precision. We will for example not standardize every occurrence of feet using the term "XXX feet", but the system will verify if the value found is lower than 300 feet, or higher than 1000 feet or between the two and this measure will then

respectively by harmonized under the terms “lt300 feet”, “gt1000 feet” and “XXX feet”.

- Organizing the extracted terms under operative concepts is another way to adapt the system to the aviation domain and to group terms considered to be equivalent (for example “bad weather” , “poor weather”). Unlike the processing of acronyms and quantities which are performed at the beginning of the linguistic analysis, this processing is carried out after the phrases have been extracted. This “conceptualization” consists in grouping together under a same term (i.e. the concept), a set of terms considered to be equivalent by domain experts. This conceptualization is therefore achieved with the help of expert knowledge and it remains valid only for the domain under study.

  Besides taking into account the semantic proximity of terms, conceptualization is used to group together hyponyms under their hyperonym: various species of birds (pigeon, seagull, etc.) are for example grouped under the “BIRD” concept. This grouping is done using our world knowledge as well as electronic resources organizing such knowledge (e.g. WordNet).

CATEGORISATION: LEARNING THE CORRELATIONS & SUGGESTIONS

The Method: Learning the correlations

In addition to the narrative texts which can be linguistically analyzed, incident and accident reports also contain some codification, i.e. information which is chosen in a closed list of category values. Air France uses its own taxonomy developed based on the SMS model; it is organized in several fields (as we have seen above in Figure 1) which have varying degrees of precision. That is why some fields cover a limited number of values (about ten) while others, since they give details on the event, cover a large number of different values (up to more than two hundred).

The basic principle behind the predictive system is to use a corpus of already coded occurrences to learn the correlation between the terms in the narratives and the values for each fields of the categorization. In other words, the system learns the existing associations between the words in the narratives and the coded category values. Figure 3 below illustrates this mechanism.

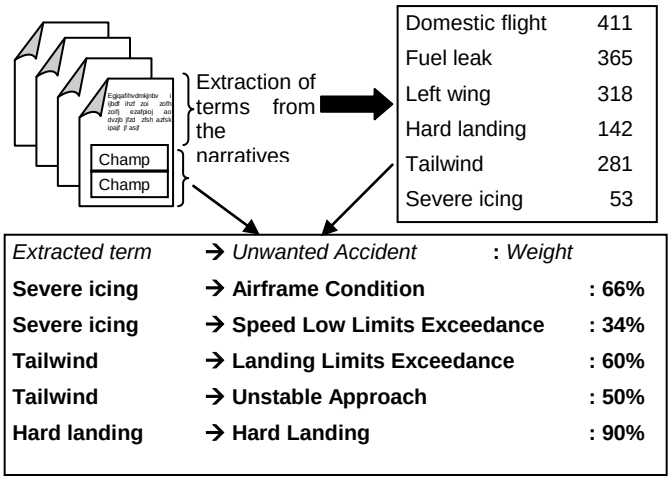

Figure 3: Learning of correlations between terms extracted from the narratives and Unwanted Accident field values

The mechanism consists in detecting how many times a given term occurs in the whole corpus on the one hand and how many times it is associated to a category value on the other hand. The probability of a narrative to be associated to a category value based on the terms which are in it can then be calculated. For example, in the case of figure 3 above, “tailwind” is associated with “Unstable Approach” in 50% of the cases.

Once these calculations have been performed, we have at our disposal a database containing all the correlation values between each term extracted from the narratives in the corpus and the associated category values. At this point, it is possible to apply a threshold to the correlation measurement. Under this threshold, the term/value association will not be taken into account because it will not be considered relevant. This correlation database is then used to suggest category values based on the terms found in new narratives.

Category value suggestion

Several terms from the same narrative can be correlated with different values but they can also be correlated with the same value. In this case, the prediction for this value is stronger. It should however be noted that the impact is not the same depending on the length of the narrative: the apparition of two terms associated with the same value will be more significant if the narrative contains about twenty words than if it contains several hundred words. Terms are even less relevant when they are repeated in the same narrative where they can co-occur with other terms which can contradict them and be linked to other values. It is not therefore here a simple arithmetic calculation (which would consist in adding the weights of each of the terms associated to the same value): the length of the narrative is taken into account to weight this sum. Once this calculation has been made, we obtain one or more values associated with a weight: the higher the weight, the more the value is

predicted to be linked to the occurrence. Figure 4 illustrates this suggestion approach[3].

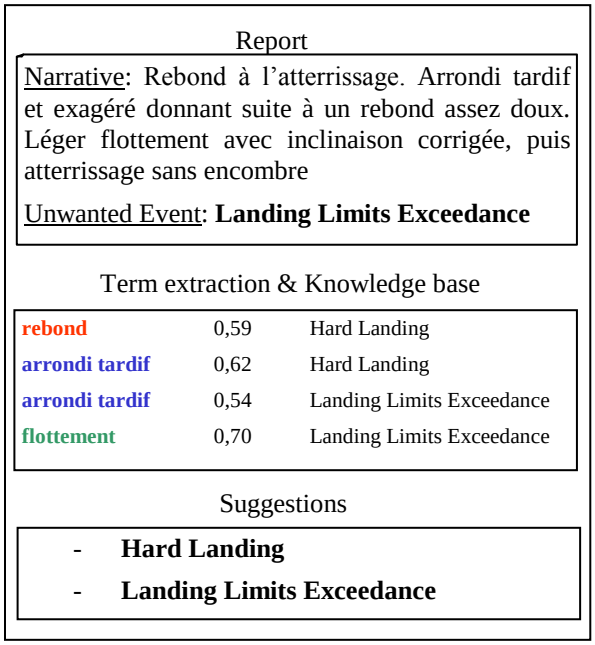

Figure 4: Suggestions

The "Term extraction & knowledge base" table contains the terms extracted from the narrative and the associations between these terms and associated category values as well as the correlation value of the term/category value pair. The associations are taken into account only if the correlation value is above a certain threshold (set here at 0.3). The terms "rebond" and "arrondi tardif" contribute to the suggestion of the "Hard Landing" value because their correlation with this value is above the 0.5 threshold. For this value to be suggested, the sum of the correlations must be above another threshold (set here at 0.9). In this example, the sum of the correlations is 1.21, allowing these two terms alone to predict the "Hard Landing" value.

Complementary approaches

Two particular processes are used to complement the suggestion method presented above.

- The method presented here sometimes encounters a problem of under-representation of data. This is particularly the case with fields for which the large number of possible values makes the calculation of relevant correlations difficult. To solve this under-representation issue, the names of the values are used. Their specificity is such that their presence in the narrative will strongly encourage the suggestion of the corresponding value for the report ("Radio congestion" for example). This amounts to using "sponsored keywords", i.e. terms whose correlation to a value will be manually reinforced. This method can also be extended for some values if necessary, i.e. if they are under-represented on the one hand and strongly specified by a term on the other hand.
- In addition, work in conjunction with domain experts has led to the definition of stable links between different types of category values. These links can be used to define incompatibilities between values which must not co-occur. For example, some threats are not coherent with certain unwanted events values (it is not possible to find a threat which takes place near a gate or an event relative to a hard landing).

Operational implementations

The learning process described above and the correlations which result from it are used to suggest the category value(s) associated with a narrative. Three major implementations emerge depending on how the suggestions are used:

- Our category value suggestion tool runs on reports used during the learning process and is used to analyze the coherence of existing coded data.
- Our tool can be used for new reports already coded in order to verify the original coding.
- Finally, our tool can be used for new reports not previously coded as an online assistance for codification, integrated in a reporting tool (such as ECCAIRS) or for a large set of reports which can then be imported back into a report database (as we do with Air France).

We can make two remarks. First, these two former implementations are used to draw the expert's attention to narratives which do not obey the general coding logic. Two cases in particular require particular attention and review by the experts: cases where the original coding and the suggested coding do not match at all and cases where no suggestions are made by the system. The first case poses the question of the quality of the original coding: the difference in coding suggests that the original coding does not obey the logic applied to the whole corpus (because the suggestion rules have been learnt on this corpus). The second case questions the quantitative and qualitative characteristics of the narrative: is it long enough and/or is the information provided clear enough to describe the event? To give an example of the second way to use our tool (on new coded reports), on a corpus of 1900 coded ASR, accident suggestions differ between original coding and suggested coding for about 26% of the cases, and the system makes no suggestion at all for 3% of the cases. In other cases, original and suggested coding coincide, reflecting the coherence of the database (71%).

Our second remark is about the particularity of the narratives from new reports (in the case of the last two implementations presented): they are in the form of raw text: they have not been analyzed by the linguistic

[3] Rebound on landing. Late and exaggerated round out after a rather soft rebound. Mild wavering with corrected inclination, then uneventful landing.

processing chain and no list of extracted terms is available. This is why a specific term searching module has been created: a Pattern-Matching module. This module searches for the terms in the correlation database (extracted during the learning process and linked to category values) in the new narratives. These terms being lemmatized forms ("crew NOTICE"), they are inflected (gender, number, conjugation) to find in the narratives all the forms in which they can appear ("the crew notices", "the crews have noticed", etc.).

Experts can use the Pattern-Matching module to gain some time while coding a corpus of reports and devote themselves to handling difficult cases for which the coding is more problematic. The Pattern-Matching module obtains good results with some values such as "Flight crew incapacitation" (correctly suggested in 85% of the cases) and can therefore be trusted with such values, leaving more time for the expert to deal with more complex occurrences.

SIMILARITY ANALYSIS

Based on the linguistic analysis and the NLP methods, we can work on the narrative parts only, without linking it to the categorization information. It is the case for the calculation of textual similarity presented below.

Limitations of categorization based strategies

Categorization based strategies, such as the one discussed above, are an essential means to augment a collection of incident reports with a coherent layer of expert analysis and a powerful tool for accessing past incident data. Nonetheless, they suffer from several inherent limitations.

Any categorization schema implies a certain compromise between ease of use and expressiveness. The more expressive and fine grained a particular schema, the more individual categories and structural complexity are needed, thus rendering the categorization process more demanding and error prone. Furthermore fine-grained schemas demand an in-depth understanding of the categories and the particular conditions when they should be used and thorough and costly training of the coders. On the other hand, a schema too simple will be maladjusted to the complexity of the physical reality it is designed to reflect.

The dynamic character of civil aviation, an ever evolving operational environment, technical innovation and new procedures imply novel and unseen risks, thus requiring a constant evolution of categorization schemas. Two issues arise. First, the procedures for introducing new categories or changing the definitions of existing ones are complicated as a consensus must be obtained within the particular circle of use of a given categorization schema. More importantly, once changes are made and an updated version of the schema is produced, there is no other way of reflecting these changes on the whole collection of reports than an extremely laborious and time consuming process of manual recategorization of past data. In reality changes are more often applied only to newly categorized data and overall coherence of the collection is lost.

Similarity score

We propose methods to automatically analyze text in order to calculate a similarity score between any two reports in a collection, by comparing their narrative parts. With no human intervention in the process, these methods produce an added layer of structure on any collection in the form of similarity links. By no means a substitute to the current categorization strategies, they provide nonetheless a complementary mean of access for safety analysts to a given collection and are insensitive to the aforementioned biases.

A similarity score is as a metric, usually in the 0 to 1 interval, measuring the degree of relatedness of the meaning of two texts. The concept is straightforward. The more two texts have in common, the higher the similarity score. A score of 1 indicates texts with identical meaning. A score of 0 indicates completely unrelated texts. The closer the score is to 1, the more related the two texts are, as is illustrated by the following short examples, extracted from the probable cause statements of NTSB[4] accident reports. The value indicated in brackets is the similarity with the first example text.

1) *"The pilot's failure to maintain directional control. A factor was the snow covered runway edge."* (sim: 1)

2) *"The pilot's failure to maintain directional control during the takeoff roll. Contributing to the accident was the snow on the runway and the snowbank."* (sim 0.87)

3) *"The pilot's failure to maintain aircraft control during the landing."* (sim 0.46)

4) *"The pilot's failure to identify a hazardous landing area. Factors in the accident are the presence of snow banks/berms on the runway, and the inadequate snow removal by airport personnel."* (sim 0.48)

In these examples we can identify two distinct factors contributing to the accidents. One is a loss of control by the pilot and the other one is snow on the runway. All four of the examples involve at least one of the two factors. Texts 1 and 2 involve both of the factors and their similarity is comparatively higher than the similarity between text 1 and texts 3 or 4 where there is only one common factor.

In order to calculate the similarity score we use techniques commonly used by search engines [3] to rank documents or web pages by order of relatedness to a given query. In a nutshell, these techniques consist in designing means to represent natural language in a way that is processable by a computer while maintaining a level of abstraction such as the representation is as independent as possible from common linguistic, stylistic and typographic variation. To

[4] National Transportation Safety Board

put it in other words, what is aimed is to represent *what* is said and not *how* it is said.

Use of similarity analysis

Simple query

The most straightforward way to exploit the automatically constructed similarity links is, given a particular report, to query the database, essentially asking the question "is this something that we have seen before?", and is particularly useful when dealing with either complex issues, implicating multiple factors, or with highly specific issues, unrepresented in a given categorization schema. The following examples are two incident reports, coming from the ASN[5] public database, which have been linked by the system:

1) *"During passengers boarding in a [MAKE/MODEL] aircraft at [AIRPORT] airport, a child felt down to the ground from the top of left forward airstairs, which this type of aircraft is equipped with. At the moment of the incident, the child was in her father´s arms, falling down straight to the ground over the airstairs´s banister. The child suffered a broken arm in the event. [...]."*

2) *"On [DATE], while boarding the airplane, a child passenger fell off the airstair to a [MAKE/MODEL] aircraft, registration [REG], at [AIRPORT] airport. The passenger was seriously injured."*

These examples illustrate two virtually identical occurrences of a particularly rare type of incident. The common factors, such as the same make and model of the aircraft and the victim being a small child, may be a clue, demanding further investigation.

Chronological plotting

We have integrated the similarity analysis into a tool which, given a report, performs chronological plotting of similar reports and allows to visually explore their temporal distributions. Figure 3 is a screenshot of the tool, displaying reports as points on a two-dimensional scatter plot having, as Y-value the degree of similarity and as X-value the date of the report. Each point corresponds to a report. Clicking on the points opens a new window and displays the corresponding report. A trend line below the main plot is also displayed based on the frequency of similar reports weighted by their corresponding similarity score. This particular example concerns volcanic ash related incidents. Several peaks of similar issues can be identified, notably a cluster in the spring of 2010, corresponding to the Iceland eruption.

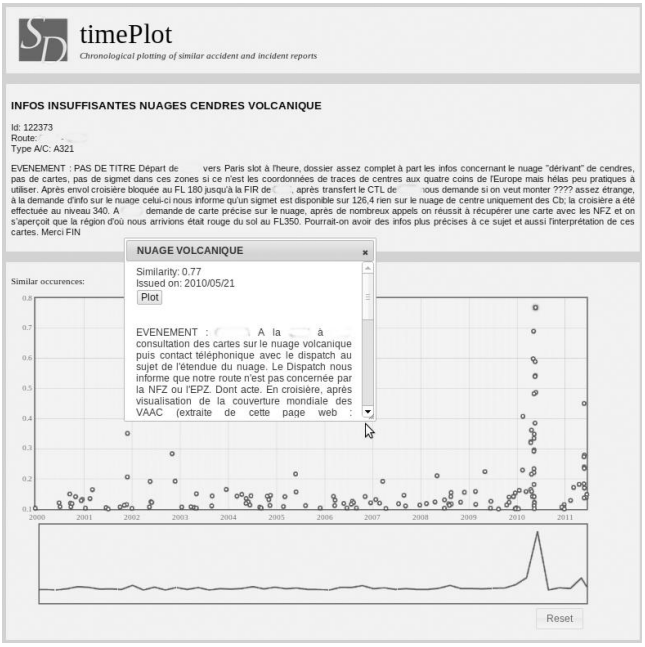

Figure 5: Screenshot of the chronological plotting tool

Proactive identification of novel risks

The above-mentioned examples illustrate the use the similarity links, by selecting a particular report of interest and exploring the similar reports and their temporal distribution. However the layer of similarity links can also be exploited in an unsupervised fashion in order to identify novel risks, by using standard data-mining techniques such as clustering.

Clustering algorithms are used to analyze a data set and produce groupings of data points based on a distance measure[6] (see [4] for details). These groupings, or clusters, represent essentially points that are close and share some characteristics.

A textbook example of such a situation is recorded in an Air France database of ASR10 reports. In early 2007, the company introduced the Runway Awareness and Advisory System (RAAS), to a part of their mid-range fleet. RAAS is a system designed to prevent runway incursions by issuing audible announcements concerning the aircraft's position while taxiing. Not long after the system was introduced, pilots started to complain that the volume of these announcements was too loud and covering their communications with ATC, thus creating a potentially dangerous situation, where a crew might miss a clearance. A new threat-category, "RAAS", was later added to the company's categorization schema.

1) "XXX équipé du système RAAS. Au départ de XXX piste 04R, ce système génère une annonce « Approaching RWY 04R » bien trop forte au moment de la clairance

[5] Aviation Safety Network

[6] A distance is the proximity of points in a given space. For our purposes we consider the inverse of the similarity measure as a distance measure.

d'autorisation décollage du contrôle, créant le risque d'une mauvaise compréhension de cette clairance et d'une incursion de piste."[7]

2) "L'annonce RAAS d'alignement a eu lieu en même temps que l'autorisation de décollage et avons eu du mal à entendre l'amendement de clairance"[8]

The similarity analysis has established links between these reports, due to the specific shared vocabulary. A distribution of RAAS-related events can be seen on the chronological plot on figure 4, showing data from 2000 up to 2011.

Combining these links with relative chronological information (how close in time reports are filed), produces input for a clustering algorithm and allows a system to isolate the set of RAAS related reports from the whole collection, based on their high similarity and temporal proximity, essentially identifying a growing problem as it is being reported.

This experiment will be detailed in a forthcoming publication.

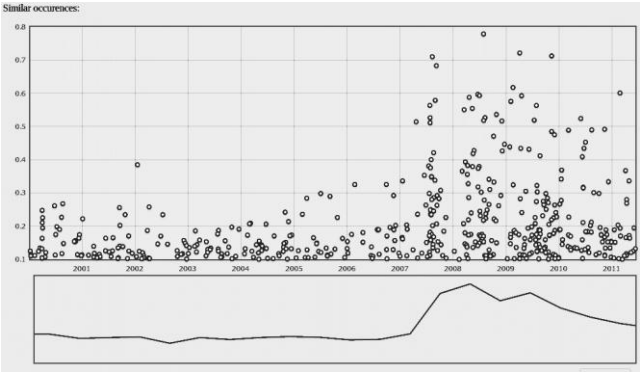

Figure 6: Distribution of RAAS related reports

CONCLUSIONS & PERSPECTIVES

Generally, we would underline the significant addition of NLP techniques for the processing and the analysis of incident and accident reports since it provides aid tools for human experts in these tasks. It is worth noting that the tools presented here are operational in English and in French and could easily be ported to other languages. The general architecture of the system and its modules are language-independent, only the linguistic analysis module would require adaptation to a particular language.

Regarding categorization, we can notice that our approach becomes increasingly relevant as taxonomies become larger, more complex and broader in scope. Issues regarding the use and analysis of natural language corpora for categorization are generic. Many domains such as biology, law or technical knowledge management rely on the building of knowledge models based on categorization. The outcome is always linked to the coherence and the descriptive capability of the categories versus the processing needs.

As for textual similarity, we can indicate that the analysis are entirely automatic, robust and require little or no supervision and virtually no other information than the text of the reports. Combined with chronological information, they can represent a highly reactive system to identify novel risks or unusual frequencies of known risks.

We are currently testing these methods within Air France's safety management effort and will continue to refine them according to the feedback we receive. A major development we are considering is extending them to take into account multilingual data and produce relevant similarity links for texts written in different languages. Such an issue exists in numerous collections, such as Air France's database, where one can find reports both written in English and in French.

We also consider researching the notion of multidimensional similarity. Combining the similarity analysis system and the automatic classification and category suggestion system, we are investigating methods to filter certain dimensions of similarity already taken into account by the categorization schema in order to isolate only those dimensions of relatedness that span across multiple categories.

REFERENCES

1. ICAO, éd. ADREP 2000 taxonomy. 2006. http://legacy.icao.int/anb/aig/Taxonomy/R4LDICAO.pdf
2. Ponvert, M. *Définition des besoins nécessaires à la mise en place d'un Data Warehouse dans le cadre du déploiment du SGS Air France*. ENAC, 2009.
3. Manning, C., Raghavan, P., et Schütze, H. *Introduction to information retrieval.* Cambridge University Press, Cambridge, 2008.
4. Witten, I., Hall, M., et Frank, E. *Data mining : practical machine learning tools and techniques.* Morgan Kaufmann, Burlington MA, 2011.
5. Bourigault D., Aussenac-Gilles N., Charlet J., Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas, Revue d'Intelligence Artificielle (RIA), " Techniques Informatiques et structuration de terminologiques, Pierrel J.-M. et Slodzian M. (Ed.), Paris: Hermès. Vol. 18, n°1/2004, pp 87-110, 2004.

[7] XXX equipped with the RAAS system. Departing from XXX runway 04R, the system produces a announcement "Approaching RWY 04R" too loud and at the same time as control take-off clearance, creating a risk of misunderstanding of that clearance and a runway incursion.

[8] RAAS alignment announcement comes at the same time as the take-off clearance and we had trouble hearing the clearance".