Computers in Industry 78 (2016) 80–95

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

Natural language processing for aviation safety reports: From classification to interactive analysis

Ludovic Tanguy [a,*], Nikola Tulechki [a,b], Assaf Urieli [a,b], Eric Hermann [b], Céline Raynal [b]

[a] CLLE-ERSS: CNRS & University of Toulouse, France
[b] CFH/Safety Data, France

ARTICLE INFO

Article history:
Received 30 November 2014
Received in revised form 18 September 2015
Accepted 18 September 2015
Available online 23 October 2015

Keywords:
Safety reports
Aviation
NLP
Document classification
Text mining

ABSTRACT

In this paper we describe the different NLP techniques designed and used in collaboration between the *CLLE-ERSS* research laboratory and the *CFH/Safety Data* company to manage and analyse aviation incident reports. These reports are written every time anything abnormal occurs during a civil air flight. Although most of them relate routine problems, they are a valuable source of information about possible sources of greater danger. These texts are written in plain language, show a wide range of linguistic variation (telegraphic style overcrowded by acronyms or standard prose) and exist in different languages, even for a single company/country (although our main focus is on English and French). In addition to their variety, their sheer quantity (e.g. 600/month for a large airline company) clearly requires the use of advanced NLP and text mining techniques in order to extract useful information from them. Although this context and objectives seem to indicate that standard NLP techniques can be applied in a straightforward manner, innovative techniques are required to handle the specifics of aviation report text and the complex classification systems. We present several tools that aim at a better access to this data (classification and information retrieval), and help aviation safety experts in their analyses (data/text mining and interactive analysis).

Some of these tools are currently in test or in use both at the national and international levels, by airline companies as well as by regulation authorities (DGAC,[1] EASA,[2] ICAO[3]).

1. Introduction

Air transportation, like other safety-critical activities, has seen the design and deployment of a large variety of safety-management procedures. Many of these efforts rely on a steady stream of reports that relate any abnormal event at any phase of activity and at any level of gravity.

This data is extremely valuable for learning lessons from past incidents and accidents, and hence for identifying new threats to safety and providing means of avoiding them. As in any complex system, the origin of these threats can be technical, organisational, environmental or human, or (most of the time) a combination of the above.

Because of this, national and international regulation bodies, as well as transport companies, store a large collection of reports for analysis.

Manual analysis of these reports is complex and requires considerable resources. Each safety event contains, in addition to other information, a description of the facts written in natural language, and each event is assigned codes from predefined taxonomies. Complexity arises, on the one hand, from the need to categorize the reports (given the size of the taxonomy, the users' knowledge, etc.) and, on the other hand, from the need to analyze and understand the reports from a global point of view. Our goal is to develop tools to help categorize and analyze the data.

CFH/Safety Data has been working on different aspects of these report systems for more than 10 years, in collaboration with the *CLLE-ERSS* linguistics laboratory. This paper is a wide-spectra presentation of the joint research we have conducted in order to integrate natural language processing (NLP) tools in the management of aviation safety reports. This work has been performed in close collaboration with both the data providers and the end users (safety experts). This article is organised as follows.

* Corresponding author.
E-mail addresses: tanguy@univ-tlse2.fr (L. Tanguy), tulechki@univ-tlse2.fr (N. Tulechki), urieli@safety-data.com (A. Urieli), hermann@safety-data.com (E. Hermann), raynal@safety-data.com (C. Raynal).

[1] Direction Générale de l'Aviation Civile.
[2] European Aviation Safety Agency.
[3] International Civil Aviation Organization.

http://dx.doi.org/10.1016/j.compind.2015.09.005

Section 2 presents a synthetic view on the existing aviation safety reporting systems and data, and summarizes the different tasks that have been identified for NLP to fulfill.

In Section 3 we present the most straightforward of our approaches: the *classification* of reports. A classical problem for NLP, it can quite easily be dealt with using supervised machine learning techniques based on textual content, and we show it succeeds when non-extreme conditions are met. Although this method is currently used by some companies and authorities, this solution is limited by the classification process itself, which is not adapted to a constantly changing environment and cannot be used for the identification of emerging threats.

We propose to address this problem with inductive methods that aim to mine patterns in the text data, and lead to the proposal of categories that can be compared to existing ones. We describe in Section 4 an experiment with *probabilistic topic models* on a large collection of reports, with mixed results that cannot conclude on the utility of this method in the specific case of extensively described and annotated data such as the repositories used in aviation safety management.

In the following two sections we present how specifically designed *interactive* tools can be useful to assist experts in their exploration of these huge and complex databases.

In Section 5 we propose a method based on the notion of document content similarity. The *timePlot* search tool is already used by several safety experts in France and enables them to quickly identify reports that are similar to a target occurrence and thus to find possible antecedents to a single event.

The last approach (Section 6) uses an *active learning* procedure in order to assist an expert circumscribe a known but not thoroughly defined aspect of incidents. Contrary to a fully inductive statistical approach, it is based on the ability of an expert analyst to quickly define the raw contours of a target category of threat. We present a proof of concept of this method that encourages us to propose such a solution to safety analysts.

In the conclusion, we discuss the extension of these techniques and processes to other fields of activity, and address the delicate problem of evaluating such techniques.

2. Overview of aviation safety reports

In 2012 the probability of dying on a single flight on one of the top 39 airlines was one in twenty million. Indeed, safety in air travel is constantly improving. [12] reports 2012 as the year with the lowest accident rate (3.2 accidents per million departures) since they started keeping the record. In the vast majority of cases, even when something serious, such as an in-flight engine malfunction, occurs, the accident is avoided and the aircraft lands safely. Even more often something could have happened but was avoided in time thanks to specific equipment, training or safety procedures.

All of these reassuring facts are the results of constant efforts at improving safety at every level of the complex system that enables air transportation. One of the procedures that helps define appropriate safety measures is incident reporting.

2.1. Principles of incident reporting

Incident reporting is a large-scale process that enables (encourages, and sometimes requires) parties to relate any abnormal event (or occurrence) to a central entity that collates and then uses this data for safety prevention purposes. This is mostly done in a non-punitive manner, i.e. the purpose is not to blame the person making the report, even if he admits that he made a mistake at some point. Quite on the contrary, such feedback loops help the personnel feel directly involved in the safety process. It should be noted that in some case parties are also invited to share their positive experience given that this kind of feedback (adequate procedure, team work, etc.) is as important as problematic events to improve safety.

Ref. [15] identifies several arguments for setting up such a procedure, the main ones being:

- incident reports indicate why an accident did not occur, and help identify both the sources of danger and the safeguards;
- incidents are much more frequent than accidents, and can be submitted to quantitative analyses, giving insights into the main sources of danger;
- the data obtained is cheap—much cheaper than the cost of accidents, especially in the industrial and transportation sectors.

In addition to these obvious advantages, regulatory decisions may compel civil aviation companies or administrations to set up a reporting system. Indeed, in most countries, reporting serious incidents to the regulation authority is mandatory.

The exact architecture of a report system varies from simple centralised repositories to complex control procedures and feedback loops, but the minimal structure is as follows:

1. The reporter writes a relatively free-form text describing the incident, along with a small set of metadata (mostly concerned with the time, the location and the equipment involved) and generally assigns a category (see below).
2. The report is checked by a receiver who assesses its compliance, and sometimes add comments, remarks and/or metadata.
3. The report is stored in a database where it is indexed according to its metadata.
4. Analysts access the database in several different ways, ranging from simple queries and statistics that estimate the frequency and evolution of incident types, to data-mining investigations in order to identify emerging dangerous situations.

As for any collection of data, organisation and indexing are crucial to its usefulness. However, the very nature of the reports' origin makes it difficult to correctly organise and index them. The spontaneity of their writing by anonymous personal (as anonymity is an important part of the non-punitive aspect of incident reporting) and the large number of reports are the two main obstacles that analysis procedures have to deal with.

2.2. Sample systems and data

Although many different reporting systems exist at different levels for companies, government agencies and NGOs, we present two of the most widely used. ASRS is a North-American database of incident reports, while ECCAIRS is a software system proposed by Europe for managing incident reports at different levels.

2.2.1. ASRS

ASRS (Aviation Safety Report System) is the oldest and most famous voluntary incident reporting program for aviation. It is managed by NASA and collects voluntarily-submitted reports of aviation events in the United States.[4] Operational since 1976, ASRS has processed over a million incident reports and averages 6736 submissions monthly (322 daily), with an increasing rate over the years. This system targets several types of events from different types of reporters: general reports from pilots, Air Traffic Control reports from controllers, maintenance reports from mechanics and cabin reports from cabin crew.

[4] http://asrs.arc.nasa.gov/.