TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Des outils de TAL en support aux experts de sûreté industrielle pour l'exploitation de bases de données de retour d'expérience

Nikola TULECHKI
CLLE-ERSS, Université de Toulouse-Le Mirail, CNRS
nikola.tulechki@univ-tlse2.fr
Conseil en Facteurs Humains
http://www.cfh-ergonomie-linguistique.com/

Résumé. Cet article présente des applications d'outils et méthodes du traitement automatique des langues (TAL) à la maîtrise du risque industriel grâce à l'analyse de données textuelles issues de volumineuses bases de retour d'expérience (REX). Il explicite d'abord le domaine de la gestion de la sûreté, ses aspects politiques et sociaux ainsi que l'activité des experts en sûreté et les besoins qu'ils expriment. Dans un deuxième temps il présente une série de techniques, comme la classification automatique de documents, le repérage de subjectivité, et le clustering, adaptées aux données REX visant à répondre à ces besoins présents et à venir, sous forme d'outils, en support à l'activité des experts.

Abstract. This article presents a series of natural language processing (NLP) techniques, applied to the domain of industrial risk management and the analysis of large collections of textual feedback data. First we describe the socio-political aspects of the risk mangement domain, the activity of the investigators working with this data. We then present present applications of NLP techniques like automatic text classification, clustering and opinion extraction, responding to different needs stated by the investigators.

Mots-clés : REX, rapport d'incident, risque, sûreté industrielle, signaux faibles, classification automatique, clustering, recherche d'information, similarité, subjectivité.

Keywords: risk management, incident report, industrial safety, weak signals, automatic classification, information retrieval, similarity, clustering, subjectivity.

1 Introduction

Dans toute industrie hautement technologique, un incident peut avoir des conséquences désastreuses, provoquer des pertes matérielles considérables, des dégâts environnementaux ou, pire, coûter des vies humaines. La complexité de chaque opération, leurs intrications et la multiplicité des facteurs différents intervenant dans le fonctionnement de ces industries rendent les risques toujours présents et amènent les acteurs (opérateurs) à développer des stratégies de gestion de la sûreté des opérations. Avoir une vision d'ensemble sur l'état du système à tout moment est crucial pour toute démarche de maîtrise du risque et, lorsqu'il est question de macrosystèmes techniques de l'échelle d'une compagnie aérienne ou pétrolière, d'une centrale nucléaire, ou encore, à un niveau supérieur, d'un *secteur d'activité* tel que le transport aérien, acquérir des informations venant du plus près des opérations devient une tâche fondamentale et très difficile. Les politiques de retour d'expérience (REX) mises en place dans les secteurs à risque témoignent de ce besoin vital. Les REX visent précisément ce recueil systématique d'information, le plus souvent sous forme de compte rendus écrits, et sa (plus ou moins) libre transmission à toute la hiérarchie organisationnelle.

Une fois recueilli, le REX doit être correctement exploité, afin d'identifier les sources de risques. Ceci est le rôle des experts en sûreté, nos principaux interlocuteurs dans le cadre de cette recherche. Leur travail consiste à analyser des événements anormaux survenus dans un secteur d'activité donné et relatés, des incidents, quasi-accidents et accidents et, en se basant sur ces événements d'émettre des recommandations adéquates, afin que ces mêmes événements ne se reproduisent plus dans le futur. Or souvent, compte tenu de l'échelle des opérations, des politiques de recueil de REX de plus en plus développées et de la multiplications des canaux de partage d'informations liés à la sûreté entre institutions, les experts se trouvent face à une quantité de données hétérogènes qui deviennent difficilement maîtrisables de façon traditionnelle (codage manuel et statistiques classiques).

De plus, actuellement nous assistons à une évolution dans le concept même de gestion de la sûreté ; les acteurs sont incités à adopter une stratégie *proactive*, autrement dit à s'affranchir de l'analyse *a posteriori* (post accidentelle) et à identifier des risques latents avant qu'ils ne mènent à un accident majeur. Cette démarche de prévention met l'accent sur l'importance des événements mineurs qui peuvent contenir des indications sur une catastrophe à venir. « On le savait. C'était dans nos bases. Nous sommes juste passés à coté » entend-on dire les experts,le plus souvent sous anonymat, à la suite d'un drame industriel.

Le but de nos recherches, associant ergonomie et traitement automatique des langues (TAL) est donc de proposer des outils permettant d'abord un accès facilité aux contenus des bases de REX et, dans un deuxième temps des méthodes automatiques d'identification de risques émergents et de précurseurs de situations à risque. Ce projet pluridisciplinaire doit donc dans un premier temps identifier les besoins précis exprimés par les experts en sûreté, expliciter le contexte dans lequel s'inscrit leur activité, notamment les flux d'information et les contraintes politiques et sociales qui lui sont associés. Dans un deuxième temps, ces besoins seront traduits en une série de propositions opérationnels, des méthodes d'analyse automatique, ainsi que des traitements et algorithmes. À terme l'aboutissement sera une série d'outils destinés à venir en support à l'analyse de bases de REX dans une perspective d'une meilleure maîtrise du risque.

Cet article est organisé comme suit : Dans un premier temps nous ferons un tour rapide sur le concept de risque industriel en nous concentrant notamment sur les dernières évolutions dans le domaine qui placent de plus en plus l'accent sur le rôle de l'organisation dans son ensemble. Parallèlement nous mentionnerons les évolutions politiques et sociales, intervenues récemment dans certains secteurs d'activité, qui ont un impact direct sur la nature de notre objet d'étude, le REX.

Dans un deuxième temps nous décrirons le travail des experts en sûreté et leur rapport avec l'information du REX. Ayant ainsi établi le contexte général nous allons nous tourner vers les sciences de gestion et le concept de *signal faible* que nous adapterons à notre problématique.

Dans la deuxième partie de cet article, nous présenterons un éventail de méthodes et techniques de TAL, que nous adaptons à notre matériau textuel et aux besoins exprimés. Ces recherches, venant tout juste de commencer sont encore pour la plupart à un stade inachevé et fortement exploratoires, stade où nous cherchons encore à valider la pertinence des techniques par rapport aux besoins des experts. Nous commencerons par l'activité la plus aboutie à ce jour - la catégorisation automatique d'évènements. Ensuite nous présenterons l'approche d'*analyse de similarité*, encore en travaux, mais dont les premiers résultats sont encourageants. Dans un troisième temps nous développerons les pistes que nous explorons actuellement visant à exploiter davantage la notion de *similarité*

en l'associant à la fois à des méthodes de détection d'anomalie afin de repérer des événements anormaux ainsi qu'à des techniques de *clustering* afin de procéder à des regroupements d'événements similaires que nous pouvons caractériser de différents manières en fonction de leur comportement dans le temps. Enfin nous explorerons un axe de recherche différent, qui consiste à effectuer des analyses linguistiques fines sur le contenu textuel afin de repérer des variations stylistiques, afin de repérer les états émotionnels des rédacteurs de ces documents.

Chacune de ces techniques fera l'objet de publications détaillées dans le futur. De plus, étant donné la nouveauté du domaine, et le manque de protocoles d'évaluation adéquates ou de standards préétablis (contrairement aux domaines « classiques », comme le RI ou EI) nous ne sommes pas encore en mesure de proposer une évaluation chiffrée dans cet article, dont la vocation est avant tout d'introduire la problématique générale de nos recherches. Une thèse est en cours depuis le mois de janvier 2011 au laboratoire CLLE-ERSS à l'Université de Toulouse 2 - Le Mirail en étroite collaboration avec la société de conseil en ergonomie industrielle « Conseil en Facteurs Humains ».

2 REX et sûreté industrielle

2.1 Fondements du REX

Aujourd'hui, il existe un consensus total sur la nécessité de tirer des leçons d'événements passés, de dysfonctionnements, accidents, incidents ou tout autres écarts au fonctionnement normal. Le REX, que nous pouvons définir comme « toute formalisation d'un événement passé » remplit ce rôle de vecteur d'informations. A une petite échelle, lorsque peu d'acteurs sont impliqués ce processus est trivial [1], mais à l'échelle d'un macrosystème technique, tel que, par exemple, l'aviation civile au niveau européen, impliquant des centaines de milliers d'individus, des centaines de compagnies aériennes et une vingtaine de gouvernements, utiliser et faire circuler l'information devient une entreprise monumentale, mais nécessaire si l'on veut prendre en considération tous les facteurs pouvant intervenir dans la gestion du risque et adopter une approche globale envers sa maîtrise. La figure n° 1, extraite de (Rasmussen, 1997) illustre la complexité d'un macrosystème technique à risque et la multitude de forces, ayant un impact sur la sûreté, impliquées à différents niveaux, allant des opérateurs interagissant avec des machines (techniciens, pilotes etc..) passant par les syndicats, le management, les différents organismes régulateurs jusqu'aux gouvernements.

Particulièrement intéressants pour nous sont les flux d'information dans cette hiérarchie. L'un, évident, est le flux « descendant » ; législation, recommandations, formations et manuels d'utilisation visent à contraindre les opérateurs à un comportement standardisé, réputé « plus sûr » afin d'améliorer le niveau global de sûreté du système.

Le flux inverse encore moins évident. Faire remonter des informations du terrain jusqu'aux instances régulatrices nécessite un ensemble de mesures, des méthodes et un cadre juridique adéquat. Le secteur aéronautique est pionnier dans cette politique du REX global, grâce à la réglementation obligeant son recueil systématique et le partage avec des instances régulatrices au niveau national, tout comme au niveau européen. Il est utile de noter le conflit majeur suscité par le REX : la tension entre sécurité et responsabilité. Afin que le REX soit efficace, on doit pouvoir faire part des erreurs commises lors des opérations. Or « avouer » une erreur remet en cause l'opérateur qui l'a commise et peut l'exposer, dans des organisations « traditionnelles » à des sanctions éventuelles. Nous laisserons de coté le débat actuel sur le statut de « l'erreur humaine » [2] vis-à-vis de la sécurité, pour dire qu'afin de réduire le silence généré par la crainte de sanctions, et améliorer la qualité du REX, de nombreuses industries ont mis en place des politiques de *non-punition* et/ou d'anonymisation afin de favoriser le REX volontaire de la part des opérateurs. Ce nouveau canal d'information est amplement utilisé dans le secteur aéronautique.

1. Prenons un exemple de tous les jours : Une famille acquiert une nouvelle friteuse (nouvel équipement). Lors de la première utilisation, le mari (opérateur), n'ayant pas lu le manuel d'utilisation, introduit brusquement les pommes de terre fraîchement coupées dans l'huile très chaude (opération). L'accident survient immédiatement. L'huile bout et s'échappe de l'appareil (événements redouté). Après avoir nettoyé sa cuisine (récupération), il fait part de son expérience aux autres membres de la famille (REX), en les incitant à ne pas introduire les pommes de terre rapidement dans l'huile trop chaude (recommandation).

2. Sinon pour dire que le consensus est qu'il n'existe pas d'activité humaine sans erreurs, la plupart du temps récupérées par l'opérateur lui même, ces collègues ou des automates.

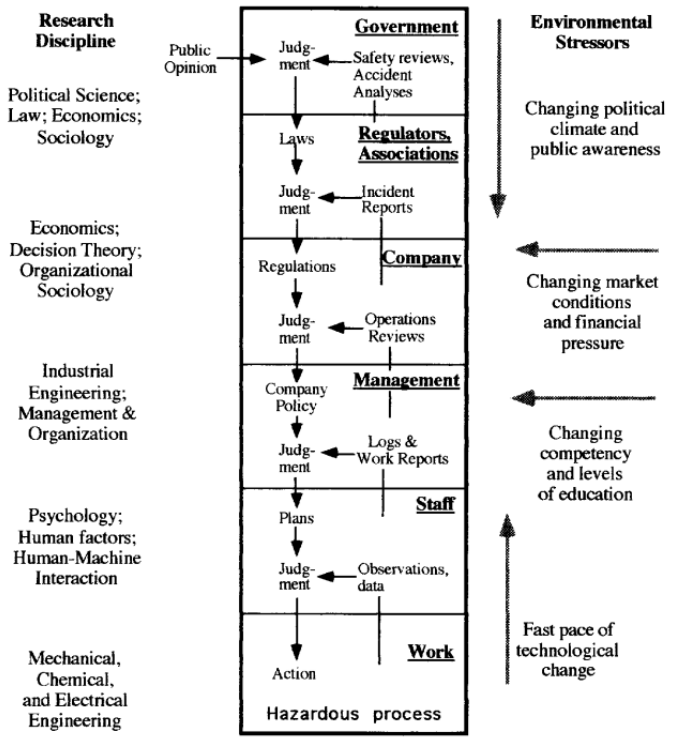

FIGURE 1 – Composantes d'une activité humaine intervenant dans le fonctionnement d'un processus dangereux

2.2 Utilisation du REX

Avoir accès aux informations ne signifie pas qu'elles seront automatiquement mises à contribution à l'amélioration de la sûreté. Nous somme rapidement arrivés à une étape, où la quantité d'information accessible dépasse les capacités d'analyse humaine. De plus puisque l'information provient le plus souvent de sources différentes, chacune avec sa propre culture vis-à-vis du recueil du REX, les informations peuvent être très hétérogènes du point de vue de leur format (bases de données, fichiers MS Excel, MS Word etc). Afin de servir de support d'analyse ces données doivent être converties en un format commun. Dans le cas de l'aviation, ou le récit d'un événement est accompagné d'un vaste ensemble de (méta)données relatives, par exemple au type d'appareil, aux conditions météorologiques, données géographiques etc, un effort de standardisation est en cours en Europe piloté par la Commission Européenne et visant à établir un tel format et un support logiciel pour son exploitation : l'environnement ECCAIRS [3] en est le fruit. Véritable « boite à outils » destinée à l'expert en sûreté, ECCAIRS propose en plus d'un format facilement échangeable, un ensemble de fonctionnalités comme un langage de requête, un navigateur spécialisé etc.

Un des étapes-clés lors de l'analyse d'un événement nouveau est sa *codification*, un procédé visant à attribuer à un évènement un ensemble de codes correspondant à ses différents facettes ; le type de l'événement (e.g sortie de piste, choc avec un oiseau, etc.) des facteurs d'environnement décrivant l'événement (conditions météo défavorables, panne d'un équipement, etc.), des facteurs contribuant à l'explication de l'événement (fatigue de l'opérateur, méconnaissance d'une procédure, etc.) En tout une liste [4] de plusieurs milliers de valeurs, organisées en une série de taxonomies, établies par l'organisation de l'aviation civile internationale (OACI), lors d'un effort de standardisation de l'analyse d'incidents en aéronautique. Une fois codifié un rapport peut être stocké dans une base et est réutilisable par la suite. Cependant puisque le contexte est en perpétuelle évolution, le schéma de codification doit être constamment mis à jour [5]. Des méthodes de classification automatique peuvent venir en aide à

3. European Coordination Centre for Accident Incident Reporting Systems : `http://eccairsportal.jrc.ec.europa.eu/`
4. Le schéma de codification ADREP actuelle est disponible à cette adresse : `http://www.icao.int/anb/aig/Taxonomy/`
5. Les perturbations du transport aérien en 2010 dues à l'éruption d'un volcan en Islande ont naturellement amenés au rajout d'une série

ces procédés de codification (voir infra).

2.3 Visée proactive, contexte dynamique et signaux faibles

Lors des investigations de véritables accidents, les experts cherchent à identifier les causes dites « primaires » de ces derniers. Il s'agit dans la plupart des cas d'une configuration particulière d'événements ou d'états, souvent clairement identifiables et signalés bien avant que l'accident lui même se produise [6].

Partant de ce constat, au cours de la dernière décennie, d'importants efforts ont été faits afin de dépasser la gestion du risque *a posteriori* et de se positionner dans une véritable démarche « proactive ». L'attention est portée non pas sur un accident qui s'est produit, mais un état des choses potentiellement dangereux, une catastrophe future que l'on peut éviter à la lumière des informations que nous avons aujourd'hui. Ainsi, les experts en sécurité sont de plus en plus sollicités pour traiter des gros volumes d'informations relatives à la sécurité, traitant de faits pouvant paraître peu importants chacun isolé et en se basant sur leur connaissance du domaine, faire des rapprochements entre ces faits et déceler des risques cachés. Parallèlement, les acteurs (pilotes, mécaniciens, opérateurs, etc.) sont incités à signaler tout évènement anormal relatif à la sécurité, ainsi qu'à s'exprimer dès qu'ils jugent qu'il y a un risque quelconque. Ce basculement vers une visée proactive dans la gestion du risque amène donc d'une part une augmentation importante du volume d'information disponible [7], puisque des faits de moins en moins éloignés de la norme prescrite sont signalés (Macrae, 2010) et d'autre part un changement du statut de cette information vis-à-vis de l'expert en sécurité, qui est amené à se concentrer sur des faits où le risque est de moins en moins explicite et, en faisant appel à son intuition et à son expériences, à chercher à aller au-delà de ce qui est réellement signalé et déceler les risques cachés.

Une telle intégration de la gestion de la sécurité dans les opérations même d'une entreprise est au cœur du modèle SGS (Systèmes de Gestion de la Sécurité) vers lequel sont amenés à s'orienter de plus en plus d'industries « à risque ». Défini comme « une façon de gérer la sécurité sous une optique commerciale » (TC, 2001), un SGS part de l'hypothèse que la sécurité au sein de l'entreprise doit devenir « l'affaire de tout le monde », amenant ainsi à une encore plus grande diversification des types d'informations relatives à la sécurité ainsi que ses sources. Ainsi l'identification d'une information importante risque de devenir de plus en plus difficile, ne serait-ce que du fait du volume et de la diversité des bases textuelles. Littéralement « noyés dans la masse » [8], l'accès à ces informations sera un générateur de frustration pour les experts en sécurité, qui doivent ne rien laisser de coté.

Les problèmes inhérents à la gestion de la sécurité dans un SGS, à savoir la diversité des sources d'information, la redondance, la nécessité absolue d'interprétation de cette information par un expert, et la disproportion entre sa fréquence et son importance, sont depuis longtemps connus des sciences de gestion. Partant d'une toute autre problématique, celle de la nécessité d'adaptation constante à un environnement commercial et concurrentiel en constante mutation, toute entreprise est amenée à mettre en place des procédés de *veille stratégique*, autrement dit d'être constamment « à l'écoute » de son environnement, pour toute information pouvant indiquer un changement futur de ce dernier.

Voulant systématiser ce processus d'écoute et d'adaptation, les chercheurs en gestion stratégique des entreprises ont forgé la notion de « signal faible » (Ansoff, 1975). Loin d'être une théorie à proprement parler, ce concept est plutôt une façon particulière et originale de voir l'information. Défini comme un « signe d'alerte précoce », le signal faible est une « information dont l'interprétation suggère qu'un événement susceptible d'être important pour l'avenir d'une firme pourrait s'amorcer » (Lesca & Blanco, 2002). L'hypothèse de base est que tout changement dans le contexte suffisamment important pour pouvoir influencer le bon fonctionnement d'une entreprise est forcement signalé bien avant qu'il ne produise des conséquences visibles par tous. De plus, dans la période

de codes en rapport avec les cendres volcaniques.

6. L'exemple de l'explosion d'une colonne dans la raffinerie de BP à Texas City au mois de mars 2005 est parlant. Ce désastre, provoquant la mort de 15 personnes, est survenu lorsqu'un nuage de vapeur, formé suite à une erreur dans la quantité de pétrole versé dans une colonne, s'est échappé et en contact avec une étincelle, s'est enflammé. Lors de l'investigation qui a suivi, six autres cas quasiment identiques, impliquant à la fois la même procédure et le même équipement, survenus au cours des dix dernières années, étaient mis au jour, aucun n'ayant tourné au cauchemar uniquement du fait de l'absence d'une source de flamme à proximité. Tous les six étaient pourtant dûment documentés, mais ce n'est que lors de l'investigation qu'un lien entre ces six occurrences est identifié.

7. Dans une grande compagnie aérienne, le nombre de nouveaux rapports d'incidents est aux alentours de 600 par mois.

8. La politique actuelle de signalement d'évènements relatifs à la sécurité actuellement mise en place préconise le signalement de *tout évènement potentiellement dangereux*. Or, on note que dans les faits, sont signalés une multitude d'évènements de routine, des « dérapages » de tous les jours, (comme par exemple des chocs avec des oiseaux au décollage pour les pilotes d'avion) qui finalement sont de peu d'intérêt pour l'expert en sécurité.

relativement longue, entre le moment du premier signalement d'un changement et le jour où ce changement devient réalité au point de menacer l'activité de l'entreprise, on peut observer une amplification de l'intensité du signalement. Inversement, la marge de manœuvre dont dispose l'entreprise diminue au fur et à mesure de cette période [9].

2.4 Caractéristiques des donnés

Les données sur lesquelles nous travaillons proviennent de bases de données différentes mises à notre disposition par des instances régulatrices de l'aviation civile, nationales et européennes, ainsi que par divers industriels dans des secteurs à risque (transports, industrie chimique, etc.). A l'heure actuelle notre corpus contient plusieurs dizaines de milliers de documents, écrits en anglais et en français et croît constamment.

La plupart des documents sont écrits dans un langage très technique, propre au secteur d'activité. Abondant d'acronymes, de termes techniques, de chiffres, de mesures, ces textes présentent, en règle générale des caractéristiques comme une variation lexicale relativement faible, peu de polysémie et une absence de constructions syntaxiques élaborés. Il s'agit de documents courts, la plupart ne dépassent pas les 500 mots.

Nous sommes en train de développer une grille de catégorisation fine de ces textes, qui sera présentée dans une prochaine publication.

3 Analyses automatiques de bases de REX

Dans cette deuxième partie nous présenterons quelques différents applications de méthodes issues du TAL aux données REX.

3.1 Catégorisation automatique d'événements

Nous avons vu que la *codification* des événements est une étape cruciale de leur analyse et permet leur réutilisabilité par la suite. Or, dans la réalité, cette tâche est effectuée de manière insatisfaisante pour plusieurs raisons. Vu la complexité des schémas de codification, contenant plusieurs centaines de classes, les codeurs attribuent souvent la classe la plus probable sans véritablement rentrer dans les détail. Les efforts de standardisation des bases de REX étant relativement récents, nous disposons de vastes quantités d'événements passés qui n'ont jamais été codés, mais qui présentent un intérêt pour des campagnes d'analyse d'aujourd'hui. Parallèlement, puisque les schémas de codification évoluent, et ce en règle générale après qu'un certain nombre d'événements de type nouveau soient survenus, de façon à justifier cette évolution, il est nécessaire de les identifier à posteriori et de leur attribuer les nouveaux codes.

Afin de répondre à ce besoin nous employons des techniques d'apprentissage automatique supervisé et plus précisément de classification automatique. Cette tâche consiste à attribuer automatiquement une classe à un individu en se basant sur les valeurs d'un ensemble de variables. Dans notre cas l'individu est un texte à classer, les variables sont les fréquences des termes dans le texte et la classe à prédire est le code de la taxonomie ADREP (voir supra). Voici le processus en détail. Nous partons d'un ensemble suffisamment large de documents déjà codifiés. Compte tenu des spécificités des textes, comme par exemple l'abondance de mesures et de noms géographiques, nous appliquerons une série de prétraitements qui réduisent ces termes à des *tokens* génériques (*mesure*, *pays*, etc). Ensuite nous procédons à une analyse morphosyntaxique en utilisant l'analyseur TreeTagger [10], suivie d'une analyse syntaxique en dépendances. Enfin en se basant sur la structure syntaxique, nous effectuons une extraction de séquences de mots en suivant les liens syntaxiques.

9. Illustrons ce propos par un exemple : prenons une entreprise spécialisée dans la fabrication de cassettes audio vierges. L'avancement de la technologie et plus précisément l'invention du CD-ROM, rend son produit obsolète et l'oblige à s'adapter en conséquence. Or le fait que le CD-ROM deviendra le support de référence est signalé bien avant que ceci ne devienne réalité. Au début, on peut imaginer des publications scientifiques qui décrivent la possibilité de stockage d'information sur un support optique. Ensuite un brevet est déposé pour ce nouveau support. Encore plus tard on commence à repérer des publications dans la presse spécialisée parlant d'un nouveau support qui vient d'être inventé, suivies de publications dans les médias généralistes, et ainsi de suite. Une progression dans la visibilité du signal est clairement perceptible et l'entreprise doit en tenir compte afin d'éviter toute surprise.

10. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html`

Ainsi la phrase :

```
Vers 200ft, déviation de l avion vers la gauche de la piste puis retour
                    sensible, à 200ft alarme autoland.
```

Est représentée par des séquences comme :

```
<mesure>, avion, gauche, piste, déviation, déviation avion, déviation
gauche, déviation avion gauche, déviation piste, déviation avion piste
```

Un classificateur est ensuite entraîné sur la base des corrélations entre les séquences extraites et les catégories à suggérer (voir (Hermann *et al.*, 2008) pour une explication an détail) qu'ont ces termes à apparaître dans un rapport codé dans une classe donnée.

Cette activité de codification automatique est, à l'heure actuelle opérationnelle dans le cadre de la collaboration de la société CFH avec les organismes régulateurs nationaux et européens.

3.2 Analyses de similarité et paramètre temporel

Une autre piste de recherche que nous avons entreprise, en utilisant de méthodes issues de la recherche d'information (RI), consiste à identifier automatiquement des événements similaires et d'étudier leur comportement dans le temps.

Les exemples ci dessus sont issus d'une base d'analyses d'accidents aéronautiques, de 1943 à nos jours, accessible au public [11]. Après un tri sur la longueur des textes, afin de ne pas inclure des rapports sans texte ou avec très peu de contenu, le corpus contient environ 14000 documents écrits en anglais.

La première étape est de calculer automatiquement un *score de similarité* pour une paire de documents donnée. En se basant sur les termes que les documents partagent, nous utilisons la similarité cosinus, métrique classique en RI pour attribuer un score compris entre 0 et 1 à chaque paire de documents dans la collection. Un score de 0 signifie une absence de termes en commun et un score de 1 - une identité complète. Ce score est obtenu en calculant le cosinus entre deux vecteurs dans un espace à *n* dimensions déterminées par le nombre de termes dans la collection. Chaque document est représenté par un vecteur en fonction des termes qu'il contient. Voici le processus en détail :

D'abord, afin de réduire la variation morphologique, nous procédons à une lemmatisation par Tree Tagger. Dans un deuxième temps nous construisons un espace termes en prenant les lemmes des noms, adjectifs et verbes contenus dans le texte. Ensuite nous construisons une matrice terme/document ou chaque ligne est un vecteur correspondant à un document de la collection dont les composants sont les importances dans son contenu de ses termes, calculées en utilisant la méthode de pondération TF/IDF (Jones, 2004). Enfin nous calculons le cosinus entre les deux vecteurs documents *A* et *B* en avec leur produit scalaire et leur norme.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

Afin d'explorer le comportement dans le temps des ensembles de documents similaires, nous avons mis en place un outil [12] qui projette ces documents sur un axe temporel. La copie d'écran ci dessous s'interprète de la manière suivante. Chaque point représente un document par rapport à un document source. les documents sont ordonnés chronologiquement sur l'abscisse et classés par similarité sur l'ordonnée. Un seuil de (arbitrairement fixé à 0.1) est appliqué afin de ne pas surcharger le graphique. Plus un point est à droite, plus il est récent et plus il est haut, plus il est similaire au document source. Ici est représenté l'ensemble d'évènements similaires au document source suivant, datant du 06/01/2003.

11. L'Aviation Safety Network, disponible à l'adresse suivante :`http://aviation-safety.net/` collecte les rapports traitant d'accidents aéronautiques sérieux.

12. Une démonstration de cet outil sur des données d'incidents aéronautiques est disponible à l'adresse suivante :`http://slow-start.org/safetyDataDemos/timePlotASN/main.cgi`

The captain's failure to attain a proper touchdown on runway, and his subsequent failure to perform a go-around, both of which resulted in a runway overrun. Factors were the company's inadequate dispatch procedures with their failure to provide all NOTAMS for the airport to the flight crew, and the snow covered runway

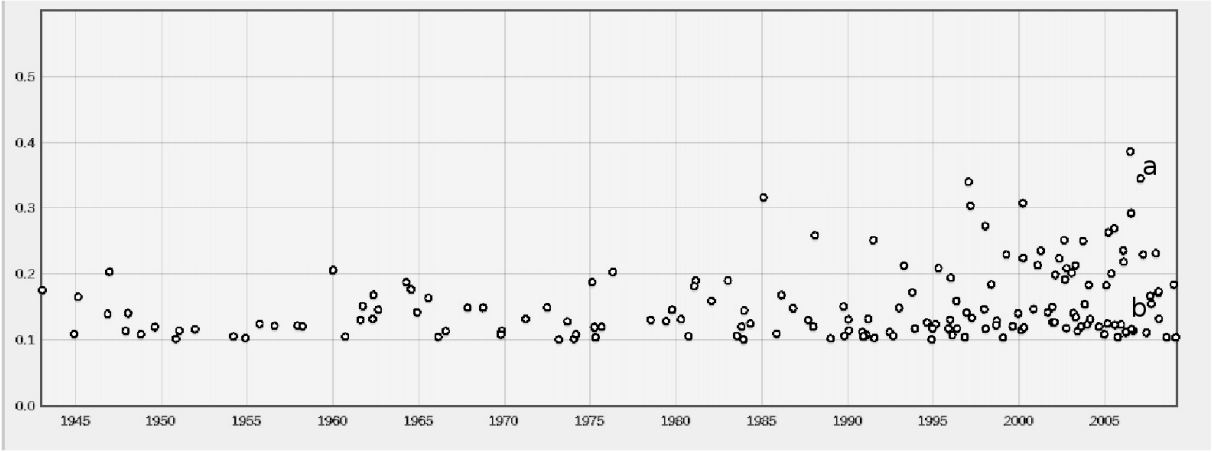

FIGURE 2 – Événements similaires sur un axe temporelle

Comparons deux documents au document source, l'un (datant du 24/01/2007, indiqué par a) relativement similaire (0.35) :

(en gras, nous avons mis les mots partagés) :

The copilot's **failure** to maintain the **proper** airspeed, and **failure** to obtain the **proper touchdown** point, and the pilot-in-command's **inadequate** supervision, which **resulted** in an **overrun.** Contributing to the accident was the PIC's **failure** to activate the speed brake upon **touchdown** and the **snow** contaminated **runway**

et le deuxième (datant du 14/09/2007, indiqué par b) un peu moins (0.15) :

The pilot's **failure** to initiate a missed **approach** and his **failure** to obtain the **proper touchdown** point while landing in the rain. Contributing to the accident were the operator's lack of standard operating **procedures** and the **inadequate** maintenance of the windshield

Nous pouvons voir comment la notion de similarité traduit un degré de ressemblance entre les deux événements ; comme le document source, les deux traitent d'atterrissage ratés, mais uniquement l'événement plus similaire mentionne des pistes enneigées.

Une entrave potentielle à l'utilisabilité de cette approche par similarité est le besoin de faire une *requête*, autrement dit de sélectionner un document,une manifestation du paradoxe de Méno [13], dont les experts en sûreté, craignant de biaiser leur analyse en faisant des présupposés, nous ont fait part. De ce fait nous envisageons de faire évoluer ces approches en employant des techniques d'apprentissage non supervisé, afin de faire émerger des regroupements naturels d'événements similaires. Ces techniques, dites de *clustering* (voir (Srivastava & Zane-Ulman, 2005) pour un exemple sur des données textuelles de type REX) permettent de s'affranchir de la requête basée sur le contenu des documents regroupés comme mode d'accès et ouvrent la voie à établir d'autres types de requêtes. Une piste que nous envisageons, peu explorée jusqu'à maintenant est celle du *profilage chronologique* (voir (Matthews *et al.*, 2010) pour un exemple d'un outil semblable destiné à des archives de presse). Une fois obtenus, les *clusters* de

13. Meno demande à Socrate comment quelqu'un peut rechercher quelque chose quand il n'a aucune idée de ce qu'est cette chose.

documents sont projetés sur un axe temporel et leur distribution chronologique est calculée. Si l'on considère les textes de l'exemple ci dessus comme les membres d'un *cluster*, ce *cluster* aurait un profil émergeant, puisque les documents récents sont beaucoup plus fréquents que les documents anciens et est susceptible d'indiquer un risque nouveau prenant de l'ampleur. A l'inverse la fréquence diminuant dans le temps d'un groupe de documents similaires, peut indiquer l'efficacité d'une recommandation nouvellement émise [14]. D'autres profils chronologiques sont aussi intéressants, comme des événements survenant à des rythmes particuliers (hebdomadaire, mensuel, annuel), par exemple.

3.3 Détection d'anomalie et événements anormaux

Les bases de REX contiennent un grand nombre d'événements similaires. Quotidiennement les avions heurtent des volatiles et ratent des atterrissages à cause d'un vent latéral. Cependant un petit nombre d'événements anormaux jamais vus jusqu'ici surviennent et il est possible de les identifier de manière entièrement automatique en utilisant des techniques de *repérage d'anomalie*, un ensemble de techniques statistiques, entièrement basées sur les données, visant à identifier dans une population les individus exceptionnels, bizarres, les *outliers* (Chandola *et al.*, 2009).

Puisqu'il s'agit de méthodes quantitatives, la principale difficulté porte sur la transformation des donnés textuelles (symboliques et qualitatives) en une série de scores numériques. Nous nous baserons pour cela sur les travaux en recherche d'information, domaine qui rencontre les mêmes difficultés. Parmi eux, certains comme (Arampatzis *et al.*, 2000) proposent des méthodes automatiques linguistiquement motivées qui visent à maîtriser la variation inhérente au langage naturel (variation lexicale, morphologique, syntaxique voire sémantique) afin d'atteindre un niveau supérieur d'abstraction, et par conséquent de produire des scores de similarité plus pertinents, scores précisément sur lesquels se basent la majorité des techniques de détection d'anomalies. Lors de quelques expériences autour du *clustering* (classification hiérarchique ascendante (CHA) (McQuitty, 1966), plus précisément), inspirés des travaux de (Ah-Pine *et al.*, 2005) nous avons pu valider l'intérêt et la faisabilité de ces techniques pour le repérage des événements « anormaux ».

Voici une esquisse de cette méthode. Partant d'une matrice de similarité, un algorithme regroupe progressivement les documents les plus similaires pour former une hiérarchie de partitions binaires incluses les unes dans les autres. Plus on monte dans la hiérarchie, plus les regroupements sont générales, plus on descend plus le critère implicite de regroupement est spécifique. Les événements anormaux, n'étant, par définition, pas similaires avec les autres, ont une tendance de former des branches hautes dans la hiérarchie (près de la racine). Nous avons expérimenté avec cette méthode utilisant un sous corpus de 110 documents, choisis manuellement afin de simuler une base avec une répartition inégale entre beaucoup de documents traitant d'événements semblables et peu de documents traitant d'événements variés. Ce corpus de test comprend 50 documents traitant des *collisions avec des oiseaux*, 50 traitant des *remises de gaz* [15] et 10 documents divers, pris au hasard. Nous avons calculé leur matrice de similarité en utilisant la méthode décrite ci-dessus avant de procéder à une CHA avec la fonction `hclust` de l'environnement d'analyse statistique *R* [16]. La figure n° 3 représente le dendrogramme produit par la CHA.

Nous nous intéresserons en particulier aux regroupements se situant haut dans la hiérarchie. Les clusters `f` et `g` contiennent la majorité des documents traitant respectivement de *collisions avec des oiseaux* et de *remises de gaz*. Plus près de la racine, les clusters `a`, `b` et `c`, contiennent les 10 documents choisies au hasard, c'est à dire les événements anormaux que nous cherchons à faire émerger. Le cluster `e` est aussi intéressant, car il contient six documents traitant à la fois de *collisions avec des oiseaux* et de *remises de gaz*, autrement dit d'événements combinant plusieurs facteurs et par ce fait intéressants pour une analyse approfondie. Cette expérience à petite échelle validant la faisabilité de la méthode, nous sommes actuellement en train d'explorer davantage cette voie en vue d'un passage à l'échelle en traitant des bases entières.

14. Nous avons récemment rencontré ce cas de figure, lors de la démonstration de ces outils aux industrielles dans une grande usine chimique. On voyait clairement les événements concernant le incidents dus aux projections reçues dans les yeux en forte baisse depuis 2007, ce qui, d'après notre interlocuteur reflétait l'effet positif de la campagne de sensibilisation au port de lunettes de protection, entreprise cette même année.

15. Une remise de gaz est une procédure d'urgence très courante lors de laquelle les pilotes décident au dernier moment d'avorter un atterrissage et de refaire un tour de l'aérodrome et une deuxième tentative d'atterrissage.

16. `http://www.r-project.org/`

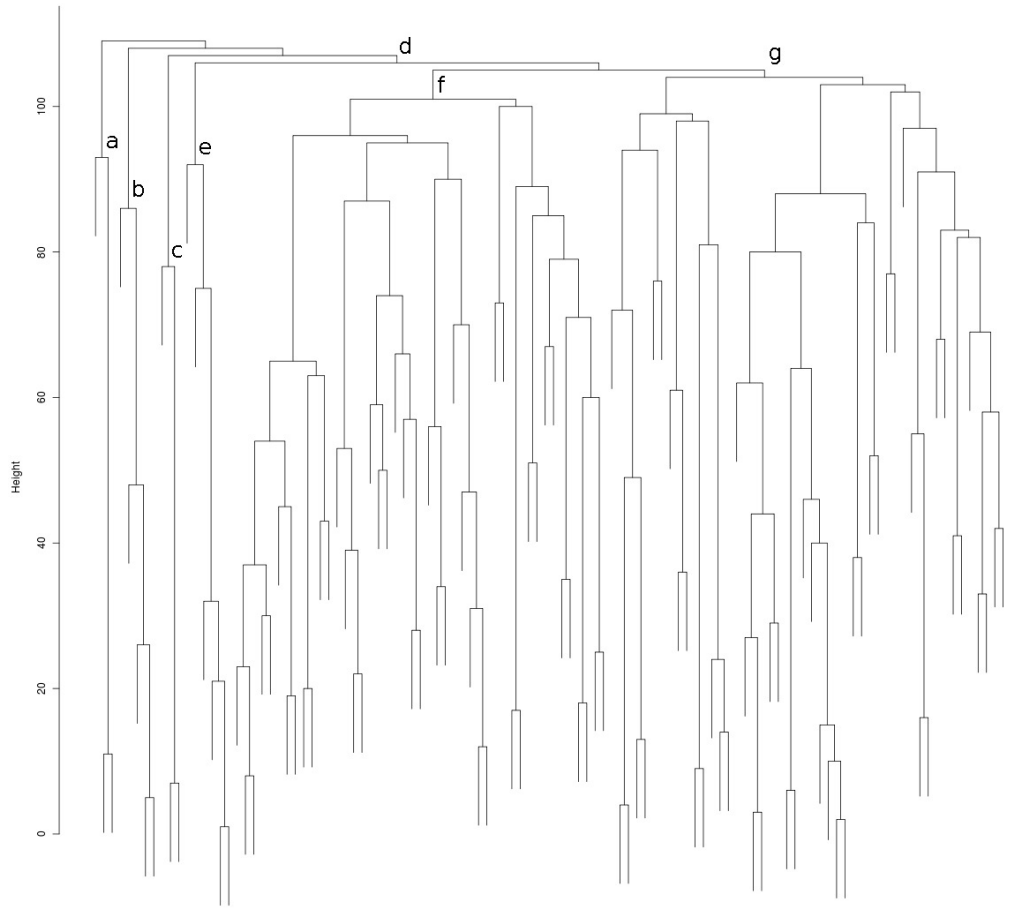

FIGURE 3 – Dendrograme des résultats de la CHA

3.4 Analyses linguistiques fines et précurseurs de situations à risque

Parallèlement aux méthodes basées sur l'ensemble des données, mentionnées ci-dessus, nous nous sommes intéressé à une autre particularité de certains donnés REX, ceux rédigés par les opérateurs eux mêmes ; le statut particulier de la place de l'auteur dans le texte. Nous avons remarqué que certains textes étaient plus « émotionnellement chargés » que d'autres. Au-delà de produire un simple récit neutre d'un événement, certains auteurs expriment des états émotionnels tels que le stress, le doute, la colère et la peur. Clairement identifiables d'un point de vue linguistique, ces états sont de véritables indicateurs de situations potentiellement à risque [17]. De plus, ayant identifié un risque récurrent et frustrés par sa non prise en compte, les acteurs manifestent souvent leur mécontentement dans leurs écrits.

Actuellement, en particulier grâce au développement du web 2.0, les travaux sur le thème du repérage automatique d'opinion et d'états émotionnels connaissent un développement spectaculaire et de nombreuses techniques innovantes voient le jour (Pang & Lee, 2008).

S'inspirant de certains travaux sur la subjectivité, nous employons une variété de traits lexicaux (adverbes axiologiques, pronoms à la première personne etc..) syntaxiques (emploi du conditionnel) et typographiques [18] que nous

17. Les pilotes, par exemple, font parfois part d'un doute ou encore d'une incompréhension d'une situation dans laquelle ils se sont trouvés. Or la maîtrise totale de la situation par les pilotes est d'une importance capitale pour la sécurité du vol et toute source de doute est à prendre au sérieux, car elle peut indiquer soit une lacune dans la formation soit un problème d'ordre organisationnel.

18. En étudiant les textes en question nous avons remarqué des pratiques tels que l'usage des majuscules ou des répétitions de point

projetons sur les textes afin de les classer par *degré de subjectivité*. Les deux textes suivants, de la même longueur et issues de la même base textuelle, illustrent cet axe, le premier faisant part de jugements personnels et écrit à la première personne, contraste fortement avec le second, beaucoup plus technique et impersonnel.

Rapport d'incident exprimant un niveau de subjectivité élevé :

LIAISON CASQUE ASSISTANT **DEFICIENTE** . [REPORT] . A l'arrivée , l'assistant est **inaudible** , **je** lui demande de changer de casque avant le départ dans 1h30 . Lors des pleins , il est toujours **inaudible** , le mécano X, présent au poste **est étonné** car à l'arrivée c'était bon . **Je** lui explique que ce n'est pas le cas et que **j'**avais déjà demandé l'échange . Liaison parfaite avec mécano X pour le litrage **faisant penser** à un autre casque . Au départ , de nouveau **inaudible** , alors que **j'**informe l'assistant de la situation , il lève l'avion (le BEACON est sur OFF **!!**) . Il demande s'il **peut pousser** . **Je** redemande un changement de casque . Attente 7 ' , rien n'est fait , avec une réception à 1/ 5 , nous poussons . A la fin **je** suis obligé de demander 3 fois qu'il se débranche et me fasse signe . **Manifestement** , la liaison est **défaillante** dans les deux sens , alors que le casque du mécano X fonctionnait parfaitement . -FIN- . [ASR] . En cas de **problème** lors du P/ B , l'équipage n'a aucune chance de comprendre ce qu'il se passe **!!! Pourquoi** cette inertie **:** on se contente de me dire « yes , ok , ... « et rien ne se passe ... **Faut -il** un **accident** pour que l'escale de Y se conforme au référentiel . Quant à la **qualité** de matériels ...

Cet exemple montre plusieurs indices reflétant un état émotionnel du rédacteur, sur lesquels nous nous basons pour classer ce document comme étant *subjectif* :
– emploi de la première personne
– emploi de constructions à verbe modal (peut pousser)
– emploi de mots évaluatifs négatifs [19] (défaillante, inaudible, problème, qualité, etc.)
– emploi de certains verbes reflétant des *états cognitifs* (penser, étonner)
– emploi de signes de ponctuation répétés (!!, ...)
Cette catégorisation unidimensionnelle, sur l'axe subjectif/objectif, n'est que la première étape de cette facette de nos recherches. Par la suite nous chercherons à établir un schéma de catégorisation plus fine et être capables de repérer séparément des états tels que le stress, le doute, mais aussi des cas de figure comme des erreurs de compréhension, de perception ou des lacunes dans les compétences mentionnés par les rédacteurs.

4 Conclusion

Nous venons de présenter nos travaux sur l'analyse automatique de bases de REX. Le domaine de la sûreté industrielle, l'analyse et l'exploitation des données textuelles issues de ces bases représentent un champ, qui, à notre connaissance, n'a jamais bénéficié de solutions utilisant des méthodes du TAL, méthodes que, nous venons de démontrer, peuvent réponde à une série de besoins exprimés dans ce secteur. De plus, nous sommes convaincus que, compte tenu de la dynamique actuelle, incitant d'un coté l'accroissement de la production de données tout autant que leur partage entre institutions les besoins d'outils spécialement adaptés se sentiront davantage.

Pour le TAL, un nombre de nouveaux défis se présentent. Un large éventail de techniques déjà connues devra être adapté à ce nouveau matériau bien particulier. Des aspects comme le flux constant de nouveaux documents et le langage très spécialisé, mais aucunement contraint, dans lequel sont rédigés la plupart d'eux, doivent être pris en compte. Enfin, le paramètre temporel, étant essentiel pour la maîtrise du risque dans un contexte dynamique, doit également occuper une place centrale dans toute approche visant à automatiser une partie de ce processus.

---

d'exclamation ou d'interrogation.

19. nous utilisons le « lexique de l'évaluation », développé par l'équipe TALN au laboratoire LINA à l'université de Nantes, que nous avons adapté à nos besoins notamment en rajoutant certains mots comme « qualité » dont nous avons vérifié le comportement axiologique dans notre corpus.

Références

AH-PINE J., LEMOINE J. & BENHADDA H. (2005). Un nouvel outil de classification non supervisée de documents pour la découverte de connaissances et la détection de signaux faibles : Rares text$^{\mathrm{tm}}$. In *Journée sur les systèmes d'information élaborés*, Île Rousse.

ANSOFF I. (1975). Managing strategic surprise by response to weak signals. *California Management Review*, **18**(2), 21–33.

ARAMPATZIS A., VAN DER WEIDE T. P., KOSTER C. H. A. & VAN BOMMEL P. (2000). An evaluation of linguistically-motivated indexing schemes. In *Proceedings of the 22nd bcs-irsg colloquium on IR research*.

CHANDOLA V., BANERJEE A. & KUMAR V. (2009). Anomaly detection : A survey. *ACM Computing Surveys (CSUR)*, **41**(3), 15.

HERMANN E., LEBLOIS S., MAZEAU M., BOURIGAULT D., FABRE C., TRAVADEL S., DURGEAT P. & NOUVEL D. (2008). Outils de Traitement Automatique des Langues appliqués aux comptes rendus d'incidents et d'accidents. In *16e Congrès de Maîtrise des Risques et de Sûreté de Fonctionnement, Avignon*.

JONES K. (2004). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, **60**(5), 493–502.

LESCA H. & BLANCO S. (2002). Contribution à la capacité d'anticipation des entreprises par la sensibilisation aux signaux faibles. In *Congres International Francophone sur les PME 6eme édition*, p. 10–1.

MACRAE C. (2010). Constructing near misses : Proximity, distance and the space between. *Risk&Regulation*.

MATTHEWS M., TOLCHINSKY P., BLANCO R., ATSERIAS J., MIKA P. & ZARAGOZA H. (2010). Searching through time in the New York Times. In *HCIR Challenge 2010*.

MCQUITTY L. (1966). Similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurement*, **26**(4), 825.

PANG B. & LEE L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, **2**(1-2), 1–135.

RASMUSSEN J. (1997). Risk management in a dynamic society : a modelling problem. *Safety science*, **27**(2-3), 183–213.

SRIVASTAVA A. N. & ZANE-ULMAN B. (2005). Discovering recurring anomalies in text reports regarding complex space systems. In *Proceedings of the 2005 IEEE Aerospace Conference*.

TC (2001). An introduction to safety management systems. *Transport Canada*.