APPORT DU TRAITEMENT AUTOMATIQUE DES LANGUES POUR LA CATEGORISATION DE RETOURS D'EXPERIENCE

NATURAL LANGUAGE PROCESSING FOR CATEGORISATION OF FEEDBACK REPORTS

Vanessa ANDREANI, Éric HERMANN et Céline RAYNAL
SAFETY DATA – CFH
13 rue Temponières 31000 Toulouse

Zakarya CHAMI et Dominique VASSEUR
EDF R&D
7 Bd Gaspard Monge 91120 Palaiseau

Résumé

La mise en place du retour d'expérience dans une entreprise implique une réflexion sur le processus global qu'il constitue : de la remontée d'information à ses objectifs en passant par les types d'analyse que l'on souhaite en faire. Cette démarche nécessite que l'on s'intéresse à la façon de structurer l'information collectée car de cette structuration va dépendre l'exploitation de la base de données de REX. Par conséquent, pour favoriser l'efficacité et la facilité d'exploration des données, on utilise généralement des « champs contraints » en vis-à-vis de champs textuels non structurés. L'expert est ainsi invité à décrire les faits en langue naturelle et également à sélectionner dans des listes prédéfinies la ou les valeurs correspondant par exemple au type d'événement. Cette catégorisation des événements n'est pas triviale, or au dire d'experts « un événement mal catégorisé est un événement perdu », et un événement perdu est un obstacle à la bonne maîtrise des risques. Afin d'aider les experts face aux taxonomies souvent lourdes et complexes à utiliser dans le travail de catégorisation, Safety Data-CFH a développé un module de catégorisation automatique dynamique et totalement autonome intégré à l'application *PLUS*. Ce module a été testé par les experts R&D d'EDF sur des fiches d'événements issus des parcs de production nucléaire et hydraulique : nous présentons les résultats obtenus.

Summary

Implementing feedback reporting in a company involves considering the global process that it represents: reporting of information, goals, or the types of analysis that experts wish to perform. This approach implies to examine the way reported information will be structured, since exploiting the feedback database will depend on this structuring. Therefore, in order to ease the efficiency and comfort of data exploration, "closed" data fields are generally used next to unstructured textual data fields. This way, experts have to describe facts in natural language, but also to select in predefined lists the appropriate value(s), for instance matching the type of event. Event categorization is not trivial, yet, as experts say, "a wrongly categorized event is a lost event", and a lost event stands in the way of proper risk management. In order to help experts facing taxonomies that can be dense and complex in their categorization task, Safety Data-CFH developed an automatic categorization module, which is dynamic and entirely autonomous, integrated to the application *PLUS*. This module was tested by EDF's R&D experts on event reports from nuclear and hydraulic power plants: we show here the achieved results.

Introduction

La gestion du risque est un domaine de recherche vaste et multidisciplinaire ; néanmoins, quelle que soit la façon dont on l'appréhende, il convient de s'appuyer sur les événements survenus dans le passé. Pour ce faire, il est nécessaire que ceux-ci aient été relatés et il est par conséquent indispensable de consigner les événements lorsqu'ils surviennent : en d'autres termes, faire du retour d'expérience (REX).
Pendant de nombreuses années, l'un des enjeux fondamentaux du REX était de collecter les informations relatives aux événements survenant sur le terrain. En effet, il s'agissait non seulement de convaincre les acteurs de première ligne de l'importance de relater les situations, mais aussi de faciliter pratiquement la façon de consigner les informations puis de les récupérer. Désormais, la collecte des REX reste encore un enjeu fort dans les entreprises, mais elle est facilitée par une culture accrue de la sécurité (« safety ») ainsi que par les avancées technologiques telle que la mise à disposition de tablettes facilitant la remontée d'informations par exemple. Par ailleurs, il est important de noter que si les premières démarches consistaient essentiellement en une collecte de données techniques à des fins d'analyse quantitative et statistique, le développement des systèmes d'information et l'intégration d'approches issues des sciences humaines dans les analyses des situations ont rendu possible l'enrichissement des analyses de sûreté avec la description de l'enchaînement des événements ou l'ajout d'éléments d'analyse et d'interprétation. Les REX se sont ainsi considérablement enrichis d'informations textuelles, rendant possible les analyses qualitatives tout en alourdissant leur exploitation étant donné leur masse conséquente.

Le volume croissant de données à analyser et leur variété a remis au premier plan du processus du retour d'expérience la question de structuration des informations recueillies. Il est effectivement crucial de porter une attention particulière à la façon d'organiser les données collectées car de cette structuration va dépendre l'exploitation qui pourra être faite de la base de REX. Pour favoriser l'efficacité et la facilité d'exploration de l'information, celle-ci est organisée en différents « champs » préétablis et on observe que des « champs structurés » ou « contraints » sont largement utilisés en vis-à-vis de champs textuels libres non structurés. De cette façon, l'expert est d'une part invité à décrire les faits en langue naturelle, et d'autre part à les classer dans des catégories prédéfinies en sélectionnant par exemple le type d'événement décrit, le matériel impliqué ou encore sa cause première (« root cause »). Cette catégorisation des événements n'est pas triviale ; or, au dire d'experts, « un événement mal catégorisé est un événement perdu », et un événement perdu est un obstacle à la bonne maîtrise des risques.

Les experts responsables de la gestion des risques étant confrontés à des bases de données textuelles, les techniques de Traitement Automatique du Langage (TAL) semblent tout indiquées pour les aider dans leur travail d'analyse des REX et plus particulièrement dans leur tâche de catégorisation. Par conséquent, Safety Data-CFH, PME spécialisée en TAL, a développé un module de catégorisation automatique afin d'apporter un appui aux experts face aux taxonomies souvent lourdes et complexes à utiliser.

Après avoir présenté les caractéristiques de la structuration des bases de données de retours d'expérience et la problématique de la catégorisation, nous expliciterons la méthode utilisée pour fournir une aide automatisée à l'expert avant de présenter la démarche grâce au projet mené par EDF et Safety Data – CFH.

Structuration et catégorisation des bases de REX

Comme nous l'avons évoqué précédemment, la collecte de retours d'expérience et le contenu des informations recueillies s'étant considérablement améliorés d'une part et les capacités de stockage n'étant aujourd'hui plus limitées d'autre part, les bases de données de REX sont de plus en plus volumineuses. Cette évolution place ces bases et leur exploitation au centre de la problématique du REX : un enjeu de taille réside dans la structuration des données car il est crucial, étant donné leur nombre croissant, de faciliter leur manipulation et les analyses qui pourront et devront en être faites.

Dans cette perspective, les informations collectées pour chaque événement sont réparties dans des champs prédéfinis, généralement hiérarchisés, qui sont de deux types. On trouve :

- Des champs de texte libre : leur contenu correspond à du texte en langue naturelle rédigé par les opérateurs de première ligne et/ou les experts en charge de l'analyse par exemple. Il s'agit en général des champs dédiés à la description des faits, l'analyse de l'événement, les mesures correctives prises, etc.
- Des champs contenant des « métadonnées », i.e. des valeurs contraintes
    - soit par leur format (des dates par exemple),
    - soit par leur contenu lorsque celui-ci correspond à une valeur prédéfinie, à choisir dans une liste existante (un établissement, un matériel, etc.).

  Ces champs contraints sont cruciaux dans la mesure où ils vont permettre non seulement d'interroger rapidement la base de données mais également d'extraire des indicateurs statistiques.

A l'intérieur du second groupe de métadonnées présentées ci-dessus (contraintes par leur contenu), on peut distinguer deux classes d'information distinctes en fonction du type de valeurs prédéfinies autorisées dans la métadonnée.

- Dans un cas, les valeurs autorisées pour un champ sont factuelles ; elles nécessitent d'être fournies par le rapporteur de l'événement lors de la remontée d'information. Il s'agit par exemple du champ dédié au site sur lequel a eu lieu l'événement ou encore le type de matériel incriminé.
- Dans l'autre cas, les valeurs prédéfinies relèvent d'une interprétation voire d'une analyse des faits. Il s'agit par exemple des champs permettant de spécifier le type d'événement ayant eu lieu ou la cause racine de la situation relatée. On parle alors de champs de catégorisation, i.e. des métadonnées dédiées à la classification thématique des événements.

La différence entre ces deux types de champ réside dans le fait que les premiers, purement factuels, doivent être fournis par une personne impliquée dans l'événement (directement ou non), tandis que les seconds peuvent être renseignés a posteriori par l'expert en charge de l'analyse. Cette possibilité est même parfois la règle : l'attribution de la cause mère de l'événement (la « root cause ») n'étant par exemple pas demandée aux opérateurs de première ligne mais réservée aux analystes.

Afin que les analyses des REX soient les plus fiables et pertinentes possibles, il est crucial que les informations regroupées dans les bases de retours d'expérience soient elles-mêmes les plus fiables possibles. Si cette nécessité requiert que la remontée d'informations suive un processus qualité, elle impose également qu'une attention particulière soit portée à la catégorisation des événements, car « un événement mal catégorisé est un événement perdu ». En effet, de la bonne « labellisation » des événements dépendent la qualité et l'exhaustivité des résultats obtenus lors d'une recherche d'événements par catégorie.
Le travail de catégorisation des événements demandé aux experts est souvent l'une des tâches obligatoires dans le traitement des REX ; il n'en demeure pas moins qu'elle peut s'avérer délicate. En effet, le choix d'une catégorie est fonction d'au moins deux paramètres. Il dépend tout d'abord de la taxonomie (i.e. les champs contraints et leurs valeurs) : celle-ci peut être complexe dans sa structuration, comporter un grand nombre de valeurs, ou des valeurs aux périmètres applicatifs proches, rendant complexe et chronophage le choix final. La détermination d'une catégorie pour un champ est également dépendante des experts métier chargés de la catégorisation des événements. En effet, la maîtrise de la taxonomie peut ne pas être la même pour tous : certains seront amenés à l'utiliser quotidiennement sur un nombre d'événements important quand d'autres n'auront à catégoriser qu'une fois par mois un petit ensemble de REX. Par ailleurs, l'opérateur humain est sujet à des biais cognitifs dans la prise de décision comme celle de catégorisation, un même expert pouvant choisir une valeur différente selon le moment où il catégorise (début ou fin de « session » de codage) par exemple.

On observe ainsi que la catégorisation est une tâche complexe qui nécessite non seulement du temps d'expert mais également une bonne connaissance de la taxonomie souvent ardue. Force est également de constater que, même si ces deux prérequis sont validés, la constance et la qualité du codage ne sont pas assurées. Face à ces constats et afin de limiter la variation de codage d'un expert à l'autre et de gagner du temps, des systèmes de catégorisation automatique ou d'aide à la catégorisation existent : nous présentons celui intégré à la plateforme d'exploration de retours d'expérience *PLUS* développée par Safety Data-CFH.

Méthode

PLUS est un environnement développé dans le but de faciliter l'exploration et l'analyse de bases de données textuelles volumineuses comme celles de REX. Disponible via une interface web, *PLUS* permet aux utilisateurs de faire des recherches fines dans les données disponibles, de comparer les descriptions des événements de la base afin de retrouver les antécédents d'une situation par exemple ou de voir la tendance d'apparition d'une problématique particulière, et enfin de catégoriser les événements contenus dans la base de données. Pour ce faire, toute base de données intégrée à *PLUS* est analysée et traitée suivant un processus établi. Ainsi, la première étape consiste à distinguer les champs textuels des métadonnées, i.e. les données libres des données contraintes, car ces deux types d'information ne nécessitent pas les mêmes traitements. En effet, les informations exprimées en langue naturelle sont par définition hétérogènes et non-structurées et tout l'enjeu de l'analyse linguistique automatique va être de les transformer en données homogènes et structurées afin de pouvoir les manipuler aussi aisément que

les métadonnées par définition structurées. Ainsi, une fois le tri opéré entre types de données, les textes libres sont analysés linguistiquement afin d'en obtenir une version épurée de toute variation de genre, de nombre, de temps mais également de forme. Une attention particulière est portée aux spécificités linguistiques des corpus analysés et un traitement dédié permet de gérer les acronymes et abréviations propres au domaine considéré. Par exemple, l'acronyme « SDC » est à rapprocher de « Salle de Commande » dans un contexte nucléaire (cf. Figure 1) tandis qu'il correspondra à « Salle des Coffres » dans le bancaire. De la même manière, la forme longue équivalente à « RG » est « Rive Gauche » dans le corpus hydraulique, alors qu'il réfèrera aux « Renseignements Généraux » dans un contexte de sécurité nationale.

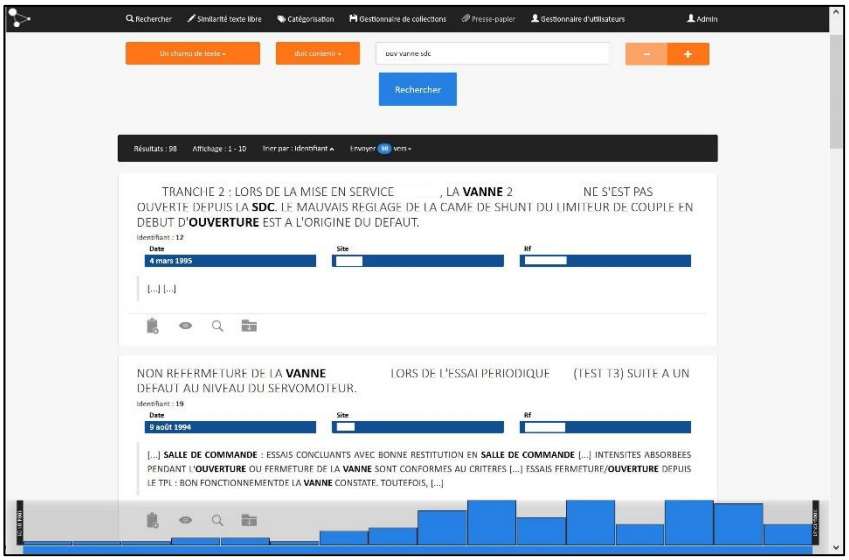

Figure 1. *PLUS* : interface de recherche

Parallèlement au traitement des données textuelles, les métadonnées sont prises en compte et analysées afin de distinguer les informations factuelles que seul un expert humain peut fournir (date, lieu de l'événement, etc.) de celles rendant compte d'une catégorisation pour lesquelles on souhaite mettre en place une aide automatisée (type d'événement par exemple). Les champs de métadonnées de ce second type vont faire l'objet d'un traitement particulier afin de fournir à l'expert un module de catégorisation automatique et totalement autonome intégré dans *PLUS*.

Pour ce faire, on utilise une technique d'apprentissage supervisé dont les descripteurs et les algorithmes d'apprentissage (basés sur les méthodes SVM[1]) sont issus de Talismane (Urieli, 2013). Le fonctionnement est le suivant : grâce à un ensemble de rapports déjà catégorisés, analysés linguistiquement et structurés via des descripteurs (qui peuvent être des termes, des groupes de mots – n-grammes – ou encore des concepts), le système va « apprendre » un modèle de catégorisation, i.e. les corrélations entre matériel linguistique et catégories, et ainsi proposer des catégories pour tous les rapports de la base de données considérée. Cet apprentissage devient « actif » grâce à l'interface de *PLUS* qui interroge les experts sur les cas « limites », et ré-entraîne le modèle tous les soirs pour prendre en compte leur avis. Ces cas correspondent aux fiches pour lesquelles la ou les suggestions proposées ne sont pas suffisamment fiables ; ils sont présentés par défaut dans l'interface de validation. Il est toutefois possible de faire apparaître les fiches pour lesquelles la catégorie suggérée a été automatiquement validée par le système si l'expert souhaite les consulter (Figure 2). Précisions qu'une suggestion est automatiquement validée uniquement lorsque le système estime que sa fiabilité est suffisamment haute. Cette estimation est propre à chaque valeur : on calcule le seuil à partir duquel la suggestion de la valeur peut être estimée fiable grâce à une validation croisée de toutes les valeurs manuellement validées. Par conséquent, une fiabilité de 90% peut être considérée suffisamment fiable pour une valeur *x* mais par pour la valeur *y* ; pratiquement dans une telle situation, les fiches d'événements pour lesquelles le système suggère la valeur *x* à 91% ne sont pas présentes par défaut dans le tableau alors que l'on voit les fiches pour lesquelles la valeur *y* est suggérée avec la même fiabilité.

[1] *SVM* : *Support Vector Machines* (Manning *et al.*, 2008)

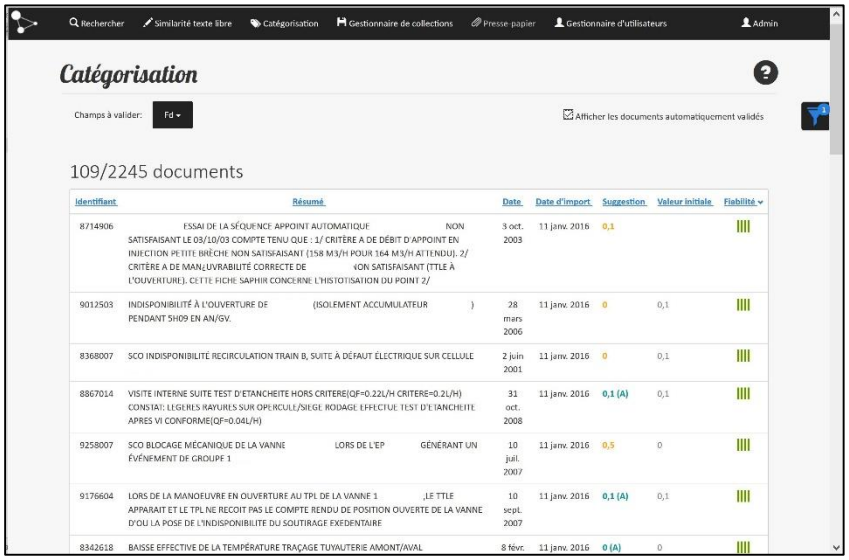

Figure 2. *PLUS* : métadonnées catégorisées et interface de validation

On a fait apparaître l'intégralité des suggestions dans la Figure 2 ci-dessus : celles qui demandent un avis expert (suggestion jaune dans la colonne « Suggestion ») comme celles automatiquement validées qui ne nécessitent pas d'intervention humaine (suggestion verte suivie de « (A) »). On va retrouver ce code couleur dans l'interface de visualisation des fiches d'événement (Figure 3) : l'expert est invité à revoir en priorité la catégorisation de la ou des métadonnées jaunes (champ « Md » ci-dessous).

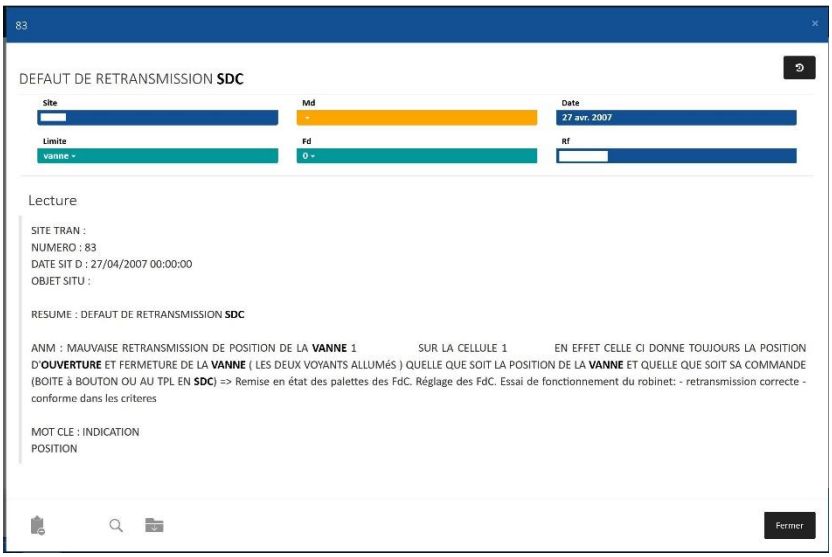

Figure 3. *PLUS* : métadonnées catégorisées et interface de validation

En cliquant sur une métadonnée faisant l'objet de la catégorisation, on obtient une nouvelle interface dédiée à la validation : l'expert peut alors confirmer ou corriger la proposition automatique (Figure 4). On voit dans la copie d'écran ci-dessous que deux catégories sont tout d'abord proposées (chacune est associée à un taux de fiabilité) dans la partie « Suggestions », mais qu'un expert pourra choisir une autre valeur si nécessaire car toutes les catégories autorisées pour le champ traité (ici Md) sont disponibles dans la partie « Taxonomie ».

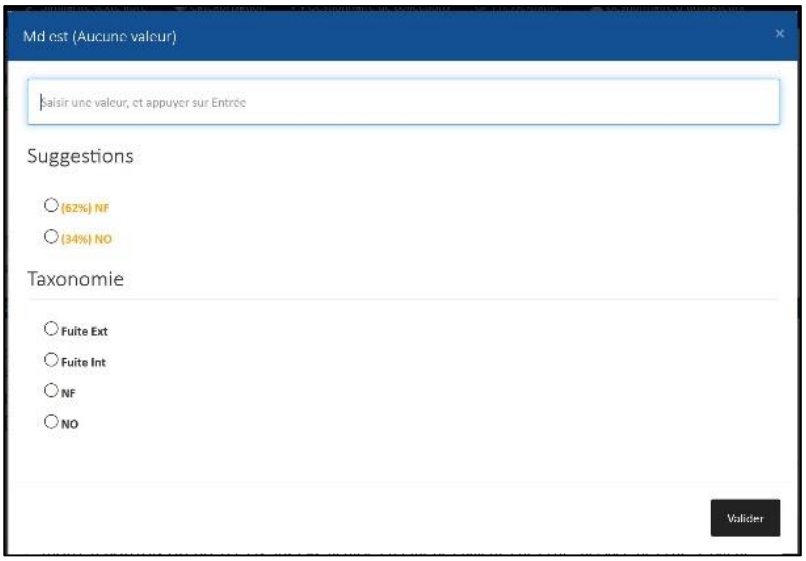

Figure 4. *PLUS* : métadonnées catégorisées et interface de validation

Insistons sur le fait que lorsqu'une suggestion est automatiquement validée par le système, l'expert pourra toujours la modifier ; elle ne lui sera seulement pas présentée en priorité afin de favoriser davantage la validation des cas moins fiables. Cela est lié au fait que toute action de l'expert est enregistrée, et que le rapport modifié est ajouté au corpus d'apprentissage et utilisé afin d'améliorer le modèle de catégorisation. Ainsi, plus l'expert donne son avis sur les « cas limites », plus le modèle de catégorisation va pouvoir s'ajuster et plus le système va s'améliorer rapidement.

La méthode décrite ci-dessus a été appliquée, à titre d'expérimentation, à des fiches d'événements issus des parcs de production nucléaire et hydraulique d'EDF ; nous en présentons les résultats à présent.

Application et résultats sur deux corpus EDF

Le processus de catégorisation automatique décrit et son intégration dans *PLUS* ont été appliqués à des fiches d'événements issus des parcs de production nucléaire et hydraulique d'EDF afin de suggérer automatiquement des catégories pour 3 champs chacun. Les deux corpus de données avaient chacun leurs spécificités[2] :

- le corpus nucléaire contenait environ 1 700 fiches composées de :
  - deux champs de texte libre décrivant l'événement,
  - deux métadonnées,
  - et les trois métadonnées à catégoriser : la Limite, le Mode de Défaillance (MD), et le Facteur de Dégradation (FD).
- le corpus hydraulique contenant environ 2 000 fiches composées de :
  - trois champs de texte libre décrivant l'événement et ses conséquences,
  - huit métadonnées,
  - et les trois métadonnées à catégoriser : l'Analyse, le Composant et le Mode de défaillance pour le REX hydraulique.

Les métadonnées à catégoriser ont été définies a priori, par les experts des domaines concernés, afin de permettre des analyses ultérieures. Par exemple, pour le corpus nucléaire, les champs Limite, Mode de défaillance et Facteur de dégradation sont utilisés pour sélectionner des événements pour un type de matériel donné et élaborer des paramètres de fiabilité tel que le taux de défaillance associé à un mode donné. Les valeurs possibles associées à chaque métadonnée sont définies soit par avis d'expert soit par des approches plus formalisées comme l'AMDE pour la définition des modes de défaillance possibles.

Comme explicité précédemment, il est important d'avoir à disposition un corpus catégorisé fiable pour permettre un bon apprentissage supervisé. Il n'est toutefois pas toujours facile de répondre à cette contrainte et on observe généralement que si des données catégorisées sont disponibles, leur fiabilité n'est pas toujours assurée ; c'est le cas ici. Pour gérer ce problème, nous avons défini un entraînement du modèle de catégorisation en deux temps : on retient d'abord uniquement les éléments « cohérents » du corpus grâce à un apprentissage croisé[3], et c'est sur ce nouveau corpus fiable que l'on va s'appuyer pour construire un modèle et l'appliquer ensuite au corpus entier.

Avant de présenter les résultats, précisons qu'ils correspondent à ceux obtenus à l'issue de la première phase de catégorisation automatique : aucune validation, confirmation ou correction d'une suggestion, n'a été faite par les experts. Par ailleurs, afin de

[2] En plus des champs explicités, deux informations sont systématiquement présentes et indispensables à la visualisation des fiches dans *PLUS* : l'identifiant unique du document et une date (celle de l'événement relaté de préférence).
[3] On divise le corpus en 10 sous-corpus ; pour chacun, on lance l'apprentissage sur les 9 autres sous-ensembles et on projette le modèle de catégorisation obtenu sur le 10e sous-ensemble : les éléments cohérents sont les fiches d'événements pour lesquels la suggestion est en cohérence avec la valeur initialement choisie par l'expert.

permettre une compréhension optimale des résultats fournis ci-après, nous explicitons les mesures utilisées dans leur évaluation, à savoir :

- La *précision* est le rapport entre le nombre de documents automatiquement et correctement attribués à la valeur *v* et le nombre total de documents automatiquement attribués à la valeur *v*. Elle permet de mettre le « bruit » en évidence : les cas où on suggère à tort une valeur.
- Le *rappel* est le rapport entre le nombre de documents automatiquement et correctement attribués à la valeur *v* et le nombre de documents initialement catégorisés avec la valeur. Elle permet de mettre le « silence » en évidence : les cas où l'on ne suggère pas la valeur attendue.
- Le *f-score* est la combinaison pondérée de deux autres mesures que sont la *précision* et le *rappel*, et permet une bonne évaluation de la fiabilité des suggestions proposées.

Ces précisions apportées, nous pouvons observer que les champs à catégoriser ont des caractéristiques spécifiques qui influencent les résultats ; nous les détaillons à présent.

- Deux de ces champs (« Limite » pour le nucléaire et « Analyse » pour l'hydraulique) ont la particularité de n'autoriser que 2 valeurs distinctes. Devant ce choix binaire, on observe tout d'abord que l'automatisation de la catégorisation fournit de très bons résultats ; on obtient par exemple un f-score global d'environ 96,5% pour le champ « Analyse » du corpus hydraulique (Table 1)[4].

	Suggestion		TOTAL V. Initiales	Précision	Rappel	f-score
	Non Retenue	***Retenue***				
Non Retenue	**94**	6	100	**94,72**	**94,14**	**94,43**
Retenue	5	**207**	212	**97,24**	**97,52**	**97,38**
TOTAL Suggestions	99	213	**312**	**96,44**	**96,44**	**96,44**

Table 1. Résultats pour le champ « Analyse » (Corpus Hydraulique)

- Cela étant, il est crucial de noter que lorsque les valeurs disponibles pour un champ sont inégalement représentées dans le corpus d'apprentissage (98% *vs* 2% pour « Limite » par exemple, cf. Table 2), les résultats s'en ressentent et la valeur très faiblement présente dans le corpus d'apprentissage est très peu suggérée par le système, le f-score associé étant par conséquent extrêmement bas (5,56% pour la valeur « HL »).

	Suggestion		TOTAL V. Initiales	Précision	Rappel	f-score
	Hors Limite	***Vanne***				
***Hors Limite* (HL)**	**3**	97	100	**50**	**2,94**	**5,56**
Vanne	3	**4856**	4859	**98,04**	**99,94**	**98,98**
TOTAL Suggestions	6	4953	**4959**	**97,98**	**97,98**	**97,98**

Table 2. Résultats pour le champ « Limite » (Corpus Nucléaire)

Cela met en évidence l'importance du corpus initial. Idéalement, on souhaite que le corpus d'apprentissage soit aussi représentatif que possible (disparités entre valeurs comprises le cas échéant) afin que le système puisse apprendre comment suggérer chacune des valeurs. Néanmoins, il n'est pas nécessaire que les valeurs soient représentées par un nombre d'occurrences identique ni même très proche pour obtenir des résultats acceptables (cf. Table 3 et Table 4), l'algorithme est prévu pour gérer la surreprésentation d'une valeur par exemple ; en revanche, une trop grande disparité dans les volumes de données risque toutefois de nuire, comme l'illustrent les résultats fournis dans la Table 2.
Notons enfin qu'une valeur absente du corpus servant à l'apprentissage ne pourra pas être suggérée par le système puisque celui-ci ne la connaîtra pas et ne saura donc pas avec quel matériel linguistique l'apparier. Toutefois, la valeur sera disponible dans l'interface de validation proposée à l'expert dans *PLUS* (cf. Figure 4) : il pourra donc la sélectionner si nécessaire. Dans une telle situation, la fiche d'événement catégorisée avec la valeur jusque-là absente du corpus d'apprentissage y sera intégrée et le nouveau calcul du modèle de catégorisation prendra en compte cette donnée : la valeur en question pourra par la suite être suggérée. Terminons en spécifiant que plus la catégorie concernée est précise – ce qui est généralement le cas des catégories initialement absentes d'un corpus d'apprentissage – moins le nombre d'occurrences requis pour l'apprentissage de la valeur doit être important, facilitant par conséquent l'intégration d'une nouvelle valeur dans le modèle.

- Les deux autres champs du corpus nucléaire (« MD » et « FD ») ont une caractéristique commune : ils acceptent chacun 4 valeurs qui sont thématiquement organisées en deux groupes de deux valeurs. Le mode de défaillance (« MD ») peut ainsi être lié à une fuite, on choisira « Fuite Ext » ou « Fuite Int. », ou à un problème d'ouverture/fermeture de la vanne, et on choisira alors entre « NO » (pour « Non Ouverture ») ou « NF » (pour « Non Fermeture »). Devant une telle organisation des catégories, on observe que les résultats obtenus sur les 4 valeurs disponibles sont relativement inégaux – le f-score va de 54% à 87% selon la valeur considérée (Table 3) – mais qu'ils s'améliorent en considérant les groupements thématiques (Table 4).

[4] Signalons que les chiffres fournis dans les tableaux ont été modifiés afin de préserver un certain degré de confidentialité ; la proportion des ensembles a en revanche été conservée afin d'illustrer correctement nos propos.

	Suggestions				TOTAL V. Initiales	Précision	Rappel	f-score
	Fuite Ext	*Fuite Int*	*NF*	*NO*				
Fuite Ext	**86**	3	8	3	100	**87,99**	**86,31**	**87,14**
Fuite Int	5	**65**	8	2	80	**88,74**	**81,35**	**84,89**
NF	4	4	**150**	25	183	**65,28**	**81,94**	**72,67**
NO	2	2	64	**57**	125	**65,68**	**45,52**	**53,78**
TOTAL Suggestions	97	74	230	87	**488**	**73,45**	**73,45**	**73,45**

Table 3. Résultats pour le champ « MD » avec les 4 valeurs disponibles

Ainsi, en ramenant les résultats aux deux valeurs « mères » Fuite ou Refus de Manœuvre, on observe que les résultats sont nettement meilleurs puisque des f-scores de 91% et 95% sont obtenus, signifiant ainsi que les « erreurs » du système sont majoritairement faites à l'intérieur d'un groupe thématique (Table 4).

	Suggestions		TOTAL V. Initiales	Précision	Rappel	f-score
	Fuite	*Refus de manœuvre*				
Fuite	**160**	20	180	**93,32**	**88,86**	**91,01**
Refus de manœuvre	12	**296**	308	**93,66**	**96,27**	**94,96**
TOTAL Suggestions	172	316	**488**	**93,54**	**93,54**	**93,54**

Table 4. Résultats pour le champ « MD » en regroupant thématiquement les valeurs

Il est intéressant de noter que ces résultats ont mis en lumière une problématique autour des périmètres des valeurs étudiées qui apparaissent alors comme sujets à discussion pour les experts métier et qu'une réorganisation des valeurs pourrait être suggérée à l'issue de cette étude.

- Enfin on observe que les deux champs complémentaires du corpus hydraulique (« Composant » et « Mode de défaillance ») cumulent les deux caractéristiques que nous venons d'évoquer. En effet, tout d'abord, les valeurs disponibles pour ces champs sont inégalement représentées dans le corpus : sur 41 valeurs disponibles pour le champ « Composant », près de la moitié (19 valeurs) sont représentées moins de 10 fois, alors que 7 d'entre elles ont entre 100 et 270 occurrences. Par ailleurs, certaines valeurs recouvrent des réalités proches et leurs périmètres respectifs sont parfois sujets à discussion.

Ces premiers résultats sont très encourageants. Les cas qui apparaissent a priori comme des « erreurs » de catégorisation doivent néanmoins être passés en revue par les experts afin de vérifier s'il s'agit effectivement de mauvaises suggestions ou si, à l'inverse, la catégorie choisie initialement par l'expert n'est pas satisfaisante et que c'est bien la suggestion qui est la valeur la plus indiquée.

Conclusion

La méthode et les résultats présentés ci-dessus permettent de montrer que l'application de la catégorisation automatique aux données de REX a plusieurs effets positifs.

- Elle permet d'une part de présélectionner, pour un champ donné, une ou plusieurs valeurs de façon systématique : elle fournit une aide à l'expert dans le choix de valeurs parmi une liste potentiellement longue tout en s'affranchissant des divers biais liés à l'activité humaine (habitude, niveau de connaissance de la taxonomie, etc.). Par là même, l'un des premiers bénéfices pour les experts en charge de catégoriser les fiches d'événement est le gain de temps que procure l'automatisation.
- Le second point à souligner concerne la qualité des données : en effet, le traitement automatisé permet une homogénéisation des pratiques de catégorisation qui influe positivement sur l'ensemble de la base de données traitées et permet, grâce à la cohérence accrue des données, une meilleure appréhension des risques.
- Cette approche automatisée permet d'autre part de s'interroger sur la taxonomie elle-même voire même de la remettre en question afin de définir aussi clairement que possible le périmètre des différentes catégories disponibles. On peut ainsi imaginer utiliser un tel module de catégorisation automatique comme outil d'aide à la définition de la taxonomie, en plus d'un outil d'aide à la catégorisation elle-même. Précisions que dans un cas où l'on redéfinirait la liste des valeurs autorisées pour un champ à catégoriser, la « reprise des données » peut se faire automatiquement.
- Cette question de la couverture de la taxonomie pose également la question de son incomplétude. En effet, l'arrivée d'un nouveau matériel dans l'entreprise pourra par exemple entraîner la nécessité d'ajouter un nouveau mode de défaillance à la liste existante. Lors de l'ajout d'une nouvelle valeur, il suffira de l'intégrer à l'interface de validation afin que l'expert puisse y accéder et ainsi, commencer à l'utiliser pour catégoriser des événements. Ceux-ci seront automatiquement intégrés au corpus d'apprentissage, permettant ainsi au système, à très court terme, de suggérer cette nouvelle valeur.

Comme précisé au préalable, les résultats présentés ci-dessus sont ceux obtenus à l'issue de la première phase de catégorisation automatique sans qu'aucune validation experte n'ait été faite. Par conséquent, la prochaine étape consiste en l'interaction homme/machine afin d'améliorer les suggestions automatiques grâce aux retours des experts sur les propositions faites jusqu'à présent. Cette étape est cruciale pour qu'un modèle de catégorisation arrive à maturité, autrement dit à un point d'équilibre où l'expert n'a plus qu'à revenir sur les rapports de REX singuliers dans leur forme ou leur contenu. En effet, cette période d'interaction entre le système et l'expert est nécessaire afin de corriger les potentiels problèmes issus de l'apprentissage initial.

Cela étant, les résultats sont suffisamment concluants en l'état pour envisager l'élargissement du périmètre traité. Cette nouvelle étape permettra de confronter le système à un ensemble de données plus nombreuses et plus variées. En effet, l'outil de catégorisation, même sans validation utilisateur dans un premier temps, apporterait quoi qu'il en soit une valeur ajoutée et une aide à l'expert grâce aux suggestions de correction, même si le modèle est d'abord imparfait. De plus, ce modèle pourrait en être amélioré, puisque le calcul s'appuierait alors sur un plus grand nombre d'éléments. On pourra également ajouter des métadonnées afin de vérifier l'intérêt de les prendre en considération de l'algorithme de catégorisation.

Références

FELDMAN R. & SANGER J., 2007, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press.
MANNING C.D., RAGHAVAN, P. & SCHÜTZE H., 2008, *Introduction to Information Retrieval*, Cambridge University Press.
MARTENS D. & PROVOST, F., 2011, « Explaining Documents' Classifications », in *Working paper CeDER*. Stern School of Business, New York University.
OLSSON F., 2009, « A literature survey of active machine learning in the context of natural language processing », in *Technical Report Swedish Institute of Computer Science.*
TELLIER I. Ed., 2009, Numéro spécial sur l'apprentissage automatique pour le TAL. *TAL*, 50(3).
TANGUY L., TULECHKI N., URIELI A., HERMANN E. & RAYNAL C., 2015, « Natural language processing for aviation safety reports: From classification to interactive analysis », in *Computers in Industry*, Elsevier.
URIELI, A., 2013. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit.* Ph.D. thesis Université de Toulouse.