UTILISATION DU TAL POUR LA CONSOLIDATION DES DONNEES DE FIABILITE ISSUES DES REX

C. Gaucher, C. Raynal, A. Urieli

UTILISATION DU TAL POUR LA CONSOLIDATION DES DONNEES DE FIABILITE ISSUES DES REX

USING NLP TO ENRICH RELIABILITY DATA FROM FEEDBACK REPORTS

Gaucher C.
EDF – DPIH/CIH
Savoie Technolac
73370 Le Bourget du Lac

Raynal C.
Safety Data – CFH
13 rue Temponières
31 000 Toulouse

Urieli A.
Joliciel Informatique
2 avenue du Cardié
09 000 Foix

Résumé

Afin d'enrichir et de consolider la base de données de fiabilité matériel et en ayant pour objectif global d'améliorer l'évaluation de la fiabilité des composants de son parc hydraulique, EDF DPIH [1] utilise l'application web *PLUS*. Celle-ci est dédiée à l'analyse de bases de données textuelles volumineuses et a pour objet d'apporter un support aux experts dans l'analyse de ces données. Ainsi, les Fiches d'Evénement et d'Exploitation rédigées par les agents EDF sont analysées par des techniques de TAL et trois champs sont catégorisés automatiquement afin, dans un premier temps, de sélectionner les fiches pertinentes, puis pour celles-ci, de sélectionner le composant et le mode de défaillance en jeu dans l'événement relaté. Nous rendons ici compte des résultats obtenus sur une base d'environ 18 000 fiches.

Summary

EDF DPIH uses the web application *PLUS* to enrich and consolidate its component reliability database, and thus to help evaluate component reliability in its hydraulic installations. PLUS uses NLP and machine learning to help experts perform data analysis on voluminous textual databases. Thus, for "event and usage reports" (FEEs) written by EDF agents, three fields are automatically categorised, first selecting the pertinent reports, and then, for these reports only, selecting the component and failure mode concerned by the event. In this article, we describe the results obtained for a database of approximately 18,000 reports.

1. Introduction

À l'instar de nombre d'entreprises, EDF collecte les retours d'expérience (REX) relatifs à l'exploitation de ses centrales. À la DPIH, les informations recueillies sont notamment utilisées pour établir des données de fiabilité pour les composants du parc hydraulique. Afin de s'assurer que les données sont fiables, les informations collectées doivent être traitées et validées par les experts. Ce travail étant coûteux en temps comme en ressources, la DPIH a décidé de poursuivre les investigations initialement menées par EDF R&D sur les bénéfices que peut permettre le Traitement Automatique des Langues (TAL) dans l'enrichissement et la consolidation de la base de données de fiabilité matériel, dans l'objectif global d'améliorer l'évaluation de la fiabilité des composants du parc. Nous présentons ici les premiers résultats d'une catégorisation automatique menée à grande échelle sur les données de la DPIH.

2. Contexte

Le traitement du REX collecté sur le parc de production hydraulique nécessite des moyens importants. À la DPIH, ce REX est notamment constitué par les FEE (Fiches d'Événements d'Exploitation) renseignées par l'exploitant. Environ 25 000 FEE sont consignées chaque année dans la base SILEX (Suivi Informatisé de L'EXploitation). Il est indispensable, pour obtenir des données de qualité, d'analyser et de valider ce corpus de données, très volumineux. À ce jour, dans le cadre du projet TAL, le corpus de travail est constitué de 18 543 FEE. Chaque FEE contient de nombreuses informations réparties dans trois types de champs :

- des champs donnant des informations factuelles telles que la date, le site où s'est produit l'événement relaté, la durée de cet événement ;
- du texte libre relatant l'événement, ses causes, ses conséquences et les mesures adoptées.

Ces deux premiers types de champs sont des informations originelles des FEE.

- des champs dont le contenu, à choisir dans une liste déroulante, relève d'une connaissance fine de l'événement.

Ce dernier type de champs est utilisé pour catégoriser les FEE. Il s'agit de définir pour chacune d'elle, si l'événement le nécessite, le composant défaillant et son mode de défaillance. Cette catégorisation est réalisée à partir de l'analyse des deux premiers types de champs : les métadonnées factuelles de l'événement et, surtout, le texte libre.

L'analyse de ce dernier, plus complexe, est absolument indispensable afin de :

- vérifier la pertinence des informations renseignées dans les autres champs ;
- compléter les données quand les champs ne sont pas remplis ;
- trouver des informations complémentaires aux champs codés ;
- catégoriser les événements en vue des analyses statistiques (pour le calcul des taux de défaillance notamment).

Cette analyse nécessitant la lecture de chaque fiche, EDF s'emploie à trouver des moyens d'alléger la charge de travail importante que représente cette tâche tout en conservant la qualité de l'analyse.
Dans cette perspective, une première étude a été réalisée par EDF R&D afin de s'assurer de l'intérêt de recourir au TAL pour réduire l'effort nécessaire à la validation des FEE (Raynal *et al.*, 2016). À l'issue de cette étude préliminaire, l'intérêt a été confirmé et le choix fait d'utiliser l'application

[1] Electricité De France - Division Production Ingénierie Hydraulique

web *PLUS* (*Processing Language Upgrades Safety*) développée par Safety Data – CFH et dédiée à l'exploitation et l'exploration de bases de données textuelles.
Le projet TAL initié à la DPIH s'appuie sur l'étude préliminaire R&D mais avec un champ d'application à la fois plus vaste et plus précis. Les principaux champs de catégorisation ont été conservés (« Analyse », « Mode de défaillance » et « Composant ») mais de nouvelles valeurs ont été attribuées pour chacun d'eux, ceci dans l'objectif de pouvoir utiliser le TAL de façon très appliquée, voire industrielle. Il a donc été nécessaire de redéfinir l'ensemble des référentiels, ainsi que le corpus d'apprentissage[2].

3. Méthode

3.1. Traitements automatiques appliqués aux données textuelles

Comme décrit précédemment, les FEE intégrées à *PLUS* contiennent des métadonnées factuelles (site, date, etc.) et un texte écrit en langue naturelle. Ces deux types de données sont traités indépendamment l'un de l'autre et les données textuelles font l'objet de traitements spécifiques. En effet, la description textuelle de toute FEE intégrée à *PLUS* fait l'objet d'une analyse linguistique automatique qui consiste en plusieurs étapes. La première[3] correspond à la réalisation d'un certain nombre de pré-traitements traitements avec la suite linguistique Talismane [4] permettant :

- de corriger les fautes d'orthographe récurrentes (« *fonctionnemnt* » est corrigé en « *fonctionnement* »),
- de standardiser les différentes formes d'un même terme telles que les abréviations (« *alim* » est équivalent à « *alimentation* »),
- d'associer les acronymes et leur version longue (« *EVC* » et « *évacuateur de crue* »).
- d'identifier des « entités nommées », à savoir des mots ou groupes de mots qui réfèrent à une entité unique : on identifie ainsi classiquement les noms de personnes, de lieux ou d'organisation comme celui des sites ou les noms de vannes.

Ces pré-traitements s'appuient à la fois sur une étude linguistique des données mais également sur des ressources fournies par EDF : en effet, si « *RG* » est l'acronyme de « *rive gauche* » dans le contexte hydraulique, il en va différemment dans le milieu de la police où l'on reconnaîtra l'acronyme des « *Renseignements généraux* ». Ainsi, l'un des objectifs des pré-traitements est de prendre en considération le plus finement possible les particularités de la langue de spécialité utilisée par EDF dans le domaine hydraulique.
Une fois ces différents pré-traitements réalisés, un module Snowball[5] est utilisé pour extraire le radical de chacun des termes employés (étape de *stemming*). De ce fait, les termes « *fonctionnement* » et « *fonctionner* » ayant la même racine « fonctionn », ils vont pouvoir être rapprochés par la suite, permettant ainsi de s'affranchir de la variation syntaxique pour accéder au contenu plus sémantique. C'est ainsi que le nombre des noms et des adjectifs (singulier *vs* pluriel) comme la conjugaison des verbes sont uniformisés. Ajoutons en guise d'exemple que, pratiquement, la recherche d'un des termes permettra de retrouver les FEE qui contiennent aussi bien le terme en question que tous ceux qui ont le même radical.

L'ensemble des traitements automatiquement effectués sur les champs textuels des FEE permet de convertir ces données initialement « non structurées » (comparativement à des métadonnées choisies dans des listes fermées par exemple) en objets structurés dans lesquels la variation linguistique est maîtrisée, permettant ainsi de réaliser des traitements statistiques pertinents et efficaces. Ce ne sont plus les mots tels qu'ils sont précisément écrits qui sont pris en considération par la suite mais leur représentation, les « traits » correspondants. Autrement dit, tous les traitements ultérieurs sont faits sur ces traits et les ensembles qu'ils constituent : ceux permettant des recherches comme ceux mis en place pour l'analyse de similarité, la création de dimensions[6] ou la catégorisation automatique.
C'est cette dernière fonctionnalité qu'il s'agit d'illustrer ici. Pour la DPIH, l'enjeu est de s'assurer de la performance de la méthode précédemment validée sur un petit corpus (étude menée avec EDF R&D) en traitant une base de données totalement brutes où la totalité de la catégorisation de trois champs spécifiques est réalisée par *PLUS* : « Analyse », « Composant » et « Mode de défaillance ».

3.2. Catégorisation des FEE

L'objectif du projet TAL est de dépouiller la base SILEX en déterminant pour chaque FEE si elle est pertinente ou non, et le cas échéant, le composant défaillant mis en jeu et son mode de défaillance. Ainsi, les FEE sont catégorisées selon trois champs principaux : « Analyse », « Composant » et « Mode de défaillance ».

- Le premier de ces champs, « Analyse », accepte seulement deux valeurs – « Retenue » et « Non retenue » – et permet aux experts de faire un premier tri entre d'une part les fiches pertinentes, nécessitant une analyse plus poussée, et d'autre part les fiches non pertinentes.
  Les fiches déclarées pertinentes sont celles faisant état de la défaillance d'un composant qui doit être identifié dans une liste préalablement établie, constituée de 118 composants. Toutes les autres fiches sont déclarées « Non retenue ». Il s'agit notamment de fiches faisant état de la défaillance d'un composant non identifié parmi les 118 connus, d'une action de maintenance sur un équipement, de la réalisation d'essais ou de mesures sur un équipement, ou encore du récit de faits divers (présence de pêcheur à l'aval d'un ouvrage, présence de voiture dans canal, etc.).
- Pour chaque fiche retenue, il convient ensuite de choisir le composant en jeu dans l'événement parmi les 118 composants existants,
- ainsi que le mode de défaillance concerné parmi les 30 possibles.

Pour permettre à un système de catégoriser automatiquement des documents, la première étape consiste à délimiter un corpus d'apprentissage, *i.e.* un ensemble de documents représentatif de chacune des valeurs disponibles pour le champ à catégoriser. Afin d'entraîner le module de catégorisation automatique de *PLUS*, une catégorisation manuelle a été réalisée par les experts de la DPIH. Il s'est agi de constituer le corpus

[2] Cette notion est précisée plus loin.
[3] La première étape est habituellement le repérage de la langue de rédaction des documents ; les FEE étant systématiquement en français, elle n'a pas lieu d'être. On fera toutefois remarquer qu'il est possible que des bases de données textuelles françaises contiennent des documents partiellement ou entièrement rédigés dans une autre langue, en anglais par exemple (c'est le cas dans le domaine aérien avec les bases de rapports de vol – *Air Safety Reports*).
[4] *Cf.* (Urieli, 2013).
[5] *Cf.* (Porter, 2001).
[6] Nous renvoyons à l'article proposé dans le cadre du λμ 21 « Repérer des dimensions dans les REX : utilisation du TAL en milieu médical » (Lagarde *et al.*, 2018) pour plus de détails sur le module de création et de gestion de dimensions disponible dans *PLUS*.

d'apprentissage nécessaire à l'initialisation du modèle de catégorisation. Ainsi, dans un premier temps 388 fiches ont été manuellement catégorisées pour le champ « Analyse », et le composant et le mode de défaillance ont été sélectionnés par l'expert pour les 108 fiches catégorisées « Retenue » ; puis dans un second temps, ces ensembles ont été complétés pour atteindre 1 581 FEE pour le champ « Analyse » et 906 FEE pour les deux autres champs ; nous y revenons en détail plus loin.

Une fois le corpus d'apprentissage délimité, l'ensemble des documents – le sous-ensemble catégorisé ainsi que les documents non catégorisés – est intégré à *PLUS* et l'entraînement se fait alors par classification supervisée probabiliste pour chacun des champs à catégoriser. On utilise une régression linéaire multinomiale[7] qui va fournir, pour chaque FEE, la probabilité d'appartenir à chacune des classes délimitées, *i.e.* à chacune des valeurs du champ concerné, afin de proposer ensuite à l'utilisateur la ou les valeurs les plus probables[8].

L'application est configurée de telle sorte que la catégorisation automatique peut être faite séquentiellement, un champ pouvant faire l'objet de la catégorisation uniquement si un autre champ, lui aussi catégorisé automatiquement, répond à une contrainte prédéfinie. C'est le cas pour la catégorisation des FEE : le système suggère d'abord une valeur pour le champ « Analyse » puis, si la valeur suggérée et validée [9] est « Retenue », des suggestions sont proposées pour les champs « Composant » et « Mode de défaillance ». Si la valeur suggérée pour « Analyse » est « Non retenue », la mention « Aucune valeur » est indiquée pour ces deux champs.

Chaque FEE fait l'objet de ce traitement et les suggestions proposées pour les 3 champs sont visibles dans l'interface de visualisation du document (Figure 1).

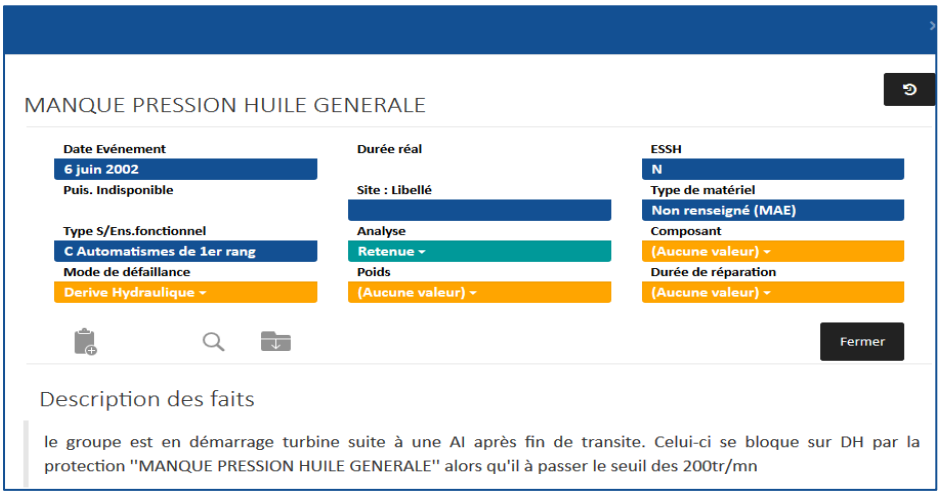

Figure 1. Visualisation d'une FEE : métadonnées factuelles et champs d'analyse

Chaque suggestion est associée à un taux de fiabilité afin de donner une indication à l'expert quant à la confiance qu'il peut avoir dans la valeur suggérée. Cette information est donnée en face de chaque valeur lorsque l'utilisateur est sur l'interface de validation, et est également signalée par une icône dans l'interface dédiée à la consultation globale (Figure 2). Cette interface permet de visualiser les résultats qui sont présentés en deux sous-ensembles (Figure 2) – que l'utilisateur peut regrouper s'il le souhaite :

- les fiches pour lesquelles la suggestion est très fiable et le système suffisamment sûr de lui pour qu'elles soient « automatiquement validées », *i.e.* elles ne requièrent pas l'avis de l'expert. Notons que le seuil permettant de considérer que la suggestion est automatiquement validée est défini en amont lors de la configuration du module pour chaque champ et qu'il peut/doit être ajusté. En effet, il est par défaut très haut afin de ne pas valider trop facilement des suggestions qui seraient des faux positifs. Nous revenons plus loin sur la méthodologie d'ajustement qui est mise en place.
- les fiches que le système préconise à l'expert de revoir car la fiabilité de la suggestion n'atteint pas le seuil requis pour que celle-ci soit validée automatiquement.

[7] LibLinear (Fan *et al,* 2008). Des expériences sont en cours pour comparer les résultats d'une régression linéaire avec ceux d'un réseau neuronal de type « Deep Learning » (LeCun *et al.* 2015).

[8] Nous renvoyons à (Tanguy *et al.* 2015) pour un état des lieux détaillé de l'approche utilisée.

[9] Une valeur suggérée peut être validée de deux façons distinctes : manuellement par l'expert, ou automatiquement par le système à la condition que sa fiabilité dépasse le seuil défini en amont. Seules les FEE dont la valeur « Retenue » a été validée selon l'une des deux options ci-dessus auront des composants et des modes de défaillance proposés.

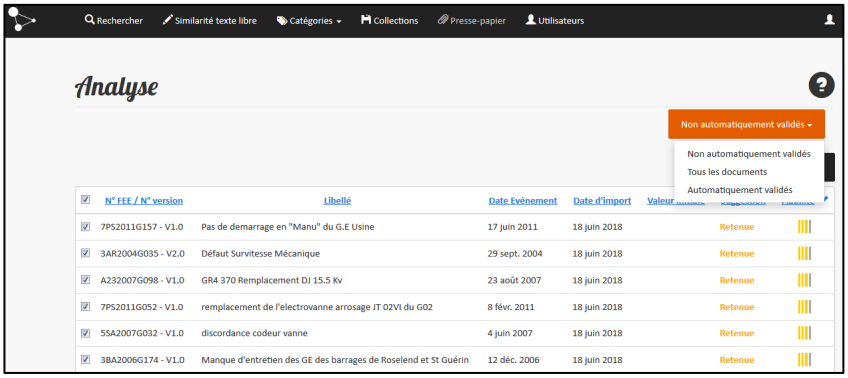

Figure 2. Tableau des résultats automatiquement validés de la catégorisation du champ « Analyse »

L'un des points clés du module de catégorisation automatique présent dans *PLUS* est qu'il est autonome et évolutif ; ainsi toute action utilisateur est prise en compte pour améliorer le modèle et donc les suggestions proposées. En effet, quotidiennement, si des validations sont faites par des utilisateurs, le système intègre ces données validées manuellement au corpus d'apprentissage pour consolider celui-ci et ainsi préciser le modèle et les suggestions proposées.
Ce travail de validation dans *PLUS* a été fait manuellement par un expert de la DPIH – et continue de l'être –, permettant de présenter en détail les résultats obtenus.

4. Résultats

Le portail *PLUS* dédié aux FEE de la DPIH est en constante évolution grâce à l'action experte de validation de la catégorisation sur les FEE qui le nécessitent ainsi que par l'ajout ponctuel mais régulier de documents. Nous présentons donc rapidement les résultats obtenus à l'ouverture du portail sur un petit ensemble de fiches (2 553 FEE) puis détaillons ceux observés pour chaque champ après avoir ajouté plusieurs milliers de documents et complété les corpus d'apprentissage.
À l'ouverture du portail *PLUS*, ce sont 2 553 fiches qui ont été intégrées ; parmi elles, un sous-ensemble était catégorisé manuellement : 308 FEE pour le champ « Analyse », soit 12% de la totalité ; parmi ces 308 FEE, 108 FEE ont été déclarées « Retenue », soit 35% des FEE du corpus d'apprentissage de « Analyse ». Ainsi, ces 108 FEE ont été catégorisées par l'expert pour les champs « Composant » et « Mode de Défaillance » afin de constituer le corpus d'apprentissage de ces deux champs.
Les résultats évalués de façon générale par échantillonnage sur le champ « Analyse » donnant satisfaction et étant équivalents aux résultats de l'étude préliminaire[10], il a été décidé d'ajouter un ensemble d'environ 16 000 FEE pour permettre d'évaluer le plus largement possible les résultats ; ainsi c'est 18 543 fiches qui sont alors accessibles dans *PLUS*. Le travail de validation du modèle de catégorisation s'est concentré dans un premier temps sur « Analyse » ; les deux autres (« Composant » et « Mode de défaillant ») sont en cours d'investigation mais nous présentons néanmoins les premiers résultats présentés dans cet article.

4.1. Champ « Analyse »

4.1.1. Résultats de la catégorisation faite par *PLUS* sur l'ensemble des FEE

Sur les 18 543 fiches désormais accessibles dans *PLUS*, 1 581 FEE (soit 8,5%) constituent à ce jour le corpus d'apprentissage de « Analyse », les autres font l'objet de la catégorisation automatique pour ce champ. Notons que le corpus d'apprentissage est relativement équilibré avec environ 57% de FEE associées à la valeur « Retenue » et 43% de FEE « Non retenue »[11]. Une fois l'apprentissage réalisé, le modèle est appliqué au corpus des FEE non préalablement catégorisées et une suggestion est proposée pour chacune d'elles. On observe alors que 86,5% de ces FEE sont automatiquement validées, et parmi elles, environ un tiers est catégorisé « Retenue » tandis que parmi les 13,5% de fiches non automatiquement validées, leur proportion représente deux tiers des FEE. Le tableau ci-dessous détaille la répartition des FEE selon la valeur suggérée la plus fiable et le degré de confiance associé qui, lorsqu'il est suffisamment haut, permet de valider automatiquement la suggestion (Table 1).

	Retenue	Non retenue
Total des FEE à catégoriser	32,9%	67,1%
FEE auto. validées : 86,5%	28,8%	71,2%
FEE non auto. validées : 13,5%	59,2%	40,8%

Table 1. Répartition des suggestions faites par *PLUS* pour le champ « Analyse » (corpus global)

4.1.2. Analyse d'un corpus test par l'expert

Les FEE intégrées à *PLUS* n'étant initialement pas catégorisées, aucune d'entre elles, a priori, ne pouvaient constituer un corpus test permettant d'évaluer la qualité des résultats. Nous l'avons par conséquent créé a posteriori : ce sont ainsi 200 FEE qui ont été choisies aléatoirement et analysées par l'expert. Parmi elles, 160 sont automatiquement validées : on observe ainsi que le corpus test est représentatif du corpus des FEE accessibles dans *PLUS* de ce point de vue. Il l'est aussi globalement concernant la répartition des suggestions entre les deux valeurs possibles comme le montrent les données détaillées dans le tableau ci-dessous (Table 2).

[10] Les résultats obtenus au cours de l'étude initiale menée avec EDF R&D avait permis d'obtenir un f-score global (sur les 2 valeurs « Retenue » et « Non retenue ») de 96% (respectivement 97,4% et 94,4%). Les mesures sont explicitées plus loin.

[11] On note la nette amélioration de la représentativité du corpus d'apprentissage puisque c'était seulement 35% des FEE du premier corpus d'apprentissage qui étaient « Retenue ».

	Retenue	Non retenue
Total des 200 FEE	33%	67%
FEE automatiquement validées : 80%	26,9%	73,1%
FEE non automatiquement validées : 20%	50%	50%

Table 2. Répartition des suggestions faites par *PLUS* pour le champ « Analyse » (corpus test)

Afin de rendre compte de la qualité des résultats, trois mesures sont généralement utilisées en TAL :

- Pour une valeur v, la *précision* est le pourcentage des FEE correctement catégorisées avec la valeur *v* parmi toutes celles auxquelles le système a attribué cette valeur. Elle permet de mettre le « bruit » en évidence : plus la précision est faible, plus les suggestions erronées sont nombreuses.
- Le *rappel* est le pourcentage des FEE correctement catégorisées avec la valeur *v* parmi toutes celles auxquelles le système aurait dû attribuer cette même valeur. Il permet de mettre le « silence » en évidence : plus le rappel est faible, plus les cas passés sous silence, *i.e.* ceux pour lesquels on ne suggère pas la valeur attendue, sont nombreux.
- Le *f-score* correspond à la moyenne harmonique de la *précision* et du *rappel*, et permet une bonne évaluation de la fiabilité des suggestions proposées.

Les résultats obtenus sur les 200 FEE du corpus test sont fournis dans le tableau ci-dessous (Table 3) ; avant d'entrer plus avant dans les détails, soulignons que l'analyse de l'expert met en évidence que 95% des FEE ont été catégorisées correctement par *PLUS* pour le champ « Analyse » (mesure d'exactitude ou *accuracy*).

	Précision	Rappel	f-score
Retenue	89,6%	95,2%	92,4%
Non retenue	97,7%	94,9%	96,3%

Table 3. Analyse du corpus test par l'expert : Résultats pour le champ « Analyse »

On observe grâce au tableau ci-dessus que la valeur « Non retenue » est très peu bruitée (précision de 97,7%), autrement dit que les suggestions de cette valeur sont rarement erronées. Le fait que la précision soit un peu moins bonne pour la valeur « Retenue », tout en conservant un rappel très satisfaisant, met en exergue le fait que le système a tendance à considérer pertinentes plus de fiches qu'il ne faudrait. Cela a un certain avantage car l'expert est ainsi assuré que très peu de fiches pertinentes seront écartées des données significatives et nécessitant un complément d'analyse. Ce comportement, certes incorrect, se traduit par un certain conservatisme des données positif pour l'expert.

4.1.3. Observations et commentaires

En regardant plus précisément les résultats selon que la valeur est automatiquement validée ou non, on peut faire quelques observations supplémentaires.

On constate tout d'abord (grâce à la Table 4 et la Table 5), que le f-score (mesure la plus globale permettant d'avoir une appréciation générale des résultats) est jusqu'à 7,4 points de pourcentage plus élevé lorsque les valeurs suggérées sont automatiquement validées et cette tendance est confirmée lorsque l'on calcule l'exactitude : elle est 96,3% pour les valeurs automatiquement validées et « seulement » de 90% lorsqu'elles ne le sont pas, confirmant ainsi la fiabilité de l'outil dans la catégorisation mais également dans son autoévaluation.

Par ailleurs, on note une précision presque « parfaite » pour la valeur « Non retenue » lorsqu'elle est automatiquement validée : si quelques fiches sont catégorisées « Retenue » à tort, la quasi intégralité des fiches catégorisées et validées automatiquement « Non retenue » le sont correctement.

Automatiquement validées (160 FEE)	Précision	Rappel	f-score
Retenue	89,4%	97,7%	93,5%
Non retenue	99,1%	95,7%	97,4%

Table 4. Analyse du corpus test par l'expert : Résultats pour les suggestions automatiquement validées du champ « Analyse »

Non automatiquement validées (40 FEE)	Précision	Rappel	f-score
Retenue	90%	90%	90%
Non retenue	90%	90%	90%

Table 5. Analyse du corpus test par l'expert : Résultats pour les suggestions non automatiquement validées du champ « Analyse »

Ces résultats sont jugés très bons, d'autant qu'une analyse qualitative des erreurs de catégorisation (soit 5% du corps test) met en exergue des cas intéressants ; on recense :

- des FEE qui font état d'une défaillance avérée mais dont le composant et/ou le mode défaillance ne sont pas identifiés dans l'étude. Aujourd'hui, il s'agit de 95% des cas d'erreur. Compléter le corpus d'apprentissage en ayant le souci que celles de ce type soit représentées permettra d'améliorer les résultats[12].
- des FEE qui font état d'une défaillance d'un composant identifié dans l'étude mais cette FEE n'est pas la FEE originelle de l'incident. Dans ce cas, seul l'expert peut avoir cette connaissance et écarter la FEE doublon afin de ne pas compter deux fois le même incident. Ce type de situation pose la question du suivi de l'incident dans la base de REX.

4.2. Champs « Composant » et « Mode de défaillance »

4.2.1. Analyse du corpus test par l'expert

Le travail d'analyse manuelle a également été réalisé sur les champs « Composants » et « Mode de défaillance », pour les FEE concernées, c'est-à-dire celles pour lesquelles la valeur suggérée et validée pour le champ « Analyse » est « Retenue », soit 42 FEE sur les 200 du corpus test (les 26,9% de la Table 2). Pour ces deux champs le corpus d'apprentissage est constitué de 906 FEE, soit les 57% de FEE catégorisées manuellement « Retenue ». Si on évalue à environ un tiers la proportion de FEE « Retenue » sur l'ensemble des FEE intégrées dans *PLUS* (ce qui correspond au 32,9% du corpus global – *cf.* Table 1 – et au 33% du corpus test – *cf.* Table 2), ce corpus d'apprentissage correspond à environ 15% des FEE qui doivent faire l'objet d'une catégorisation[13].

[12] Notons qu'en analysant ces fiches pour les intégrer au corpus d'apprentissage du champ « Analyse », l'expert les intégrera en parallèle dans les corpus d'apprentissage des deux autres champs, permettant là aussi de mieux reconnaître le composant et le mode de défaillance en jeu.

[13] On observe que ce corpus d'apprentissage est, proportionnellement, presque deux fois plus important que celui pour le champ « Analyse » (qui représente 8,5% du corpus global) ; le nombre de valeurs à considérer étant bien plus important, cela n'a non seulement rien d'étonnant mais devra également faire l'objet d'une analyse quant à la bonne

On rappelle que pour ces champs l'outil *PLUS* doit sélectionner la valeur du champ « Composant » parmi 118 valeurs possibles, et celle du champ « Mode de défaillance » sur les 30 connues. Dans les faits l'outil suggère pour chacun de ces champs 1 à 3 valeurs associées à un degré de confiance (
).

Les premiers résultats de l'analyse manuelle des 42 FEE « Retenue » sont les suivants[14] :

- Pour le champ « Mode de défaillance », une seule FEE est validée automatiquement, mais globalement, le taux de réussite est de 88%, c'est-à-dire qu'une valeur parmi les 3 suggérées par *PLUS* est la bonne ; cette proportion inclut les FEE pour lesquelles c'est la première suggestion (celle dont le taux de confiance est le plus élevé) qui est la bonne, elles représentent 78% des 42 FEE. Enfin, dans 12% des cas, aucune des valeurs suggérées n'est la bonne.
- Pour le champ « Composant », aucune des 42 FEE n'est validée automatiquement. En considérant les 3 valeurs suggérées par *PLUS*, on observe un taux de réussite de 76% (*i.e.* l'une des 3 valeurs est la bonne). Si on se limite à la suggestion la plus fiable, celle-ci est la bonne dans 69% des cas. Enfin, dans 14% des cas, aucune des valeurs suggérées n'est la bonne.

Il est important d'insister sur le fait que, dans sa configuration actuelle, l'outil ne suggère des valeurs pour « Composant » et « Mode de défaillance » que lorsqu'il est suffisamment sûr de lui quant à la valeur de « Analyse », que celle-ci est automatiquement validée (ou a été validée manuellement au préalable). Par conséquent, pour les FEE « Retenue » mais non automatiquement validées, aucun composant ni aucun mode de défaillance n'est suggéré ; or il s'agit ici d'une proportion importante de FEE : à savoir 30% de l'ensemble des suggestions « Retenue » du corpus test.

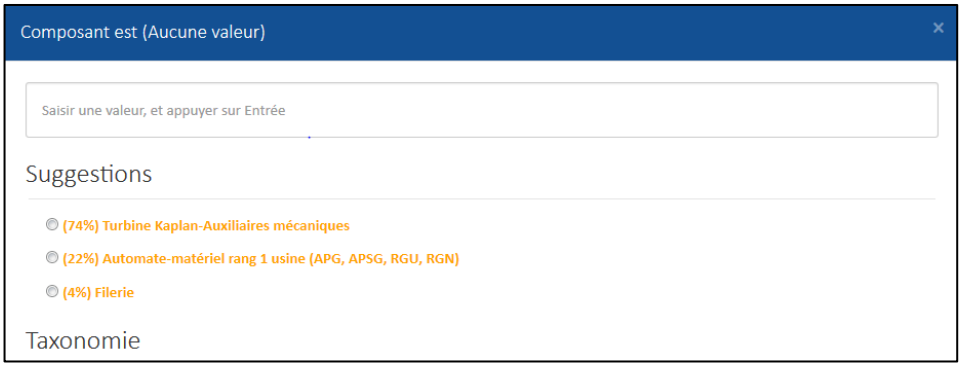

Figure 3. Visualisation des suggestions faites par l'outil pour le champ « Composant »

Or, l'analyse précédemment réalisée sur le champ « Analyse » a montré une très bonne capacité de l'outil à identifier les FEE pertinentes. Par conséquent, le seuil de validation automatique de l'outil défini par défaut pour le champ « Analyse » doit être revu à la baisse, ou la contrainte sur la validation de ce champ supprimée, ceci afin que *PLUS* suggère des valeurs pour les deux autres champs dès qu'il estime la FEE « Retenue ». En effet, 30% de FEE supplémentaires catégorisées par *PLUS*, ce sont autant de données qui permettent d'évaluer l'outil mais également d'aider l'expert lors de l'analyse des FEE dans un objectif d'apprentissage de l'outil.

4.2.2. Observations et commentaires

Au vu des possibilités de catégorisation (118 composants, 30 modes de défaillance) et des informations décrites dans les FEE, les résultats de la catégorisation des champs « Composant » et « Mode de défaillance » sont prometteurs. En effet, le volume de texte libre dans une FEE peut être important (plusieurs centaines de mots) avec une description exhaustive de l'événement dans laquelle le composant en cause à l'origine de l'incident n'est pas toujours mis en exergue. De plus, la décomposition du parc hydraulique en composants est telle que certains composants ne sont distingués que par leurs variantes technologiques (vanne segment, vanne wagon, vanne plate par exemple), chaque usine ayant ses propres caractéristiques. L'outil ne peut donc identifier le bon composant, c'est-à-dire la bonne variante technologique, qu'en associant le composant au site considéré (le libellé du site est une métadonnée). Ce travail d'apprentissage, réalisé au fil de l'eau, n'est pas encore terminé à ce jour (on compte plus de 400 sites). De même, certaines valeurs de « Composant » et « Mode de défaillance » ne font pas encore partie du corpus d'apprentissage (c'est le cas de 32 composants et 4 modes de défaillance) ; par conséquent, l'outil ne peut associer ces valeurs aux FEE qui le nécessiteraient ; rappelons que l'apprentissage étant quotidien, dès lors que les experts affecteront quelques FEE à ces valeurs, elles pourront être suggérées par *PLUS* ensuite.

Ainsi, les principales raisons des erreurs réalisées par *PLUS* sont dues à un apprentissage encore insuffisant au regard des nombreuses possibilités pour les champs « Composants » et « Mode de défaillance ». Ceci pose la question de la constitution du corpus d'apprentissage : quel nombre minimal de FEE est nécessaire, pour chaque valeur, pour un corpus d'apprentissage exhaustif ? La réponse n'est pas univoque et dépend de plusieurs critères comme la spécificité de la valeur (est-elle liée, ou non, à des problématiques très typiques telles que l'orage ou le gel par exemple pour les modes de défaillance) et la représentativité de la valeur dans l'ensemble du corpus (cette valeur est-elle fortement représentée ou non). La Table 6 met notamment en évidence la conséquence d'un déséquilibre dans la représentativité de différentes valeurs au sein du corpus. Ainsi, on observe que le mode de défaillance MD1, fortement représenté dans le corpus d'apprentissage (53,7%), est suggéré parfois à tort par l'outil (précision de 79,3%) mais qu'en revanche son rappel est

représentativité de ces FEE d'apprentissage, nous y revenons plus loin.

[14] Nous présentons ici la première analyse des résultats obtenus en termes de suggestion correctes/incorrectes sans aller plus dans le détail, c'est pourquoi nous ne donnons pas les scores de précision, rappel et f-score, et nous en tenons à « l'*accuracy* », soit la pertinence, ou exactitude, des résultats.

excellent (100%). A contrario, pour une valeur moins représentée, l'outil aura tendance à ne pas suggérer suffisamment la valeur ; c'est par exemple le cas pour le mode de défaillance MD2 qui a un rappel de 80%. Enfin, pour une valeur quasi inexistante dans le corpus d'apprentissage, MD3 avec un taux de représentativité de 1,1%, l'outil n'est pas encore en mesure de suggérer la valeur. Ces résultats mettent en évidence qu'un travail doit être effectué quant à l'homogénéisation du corpus d'apprentissage.

	Représentativité dans le corpus d'apprentissage	Précision	Rappel
MD1	53,7%	79,3%	100%
MD2	10,6%	100%	80%
MD3	1,1%	0%	0%

Table 6. Résultats pour le champ « Mode de défaillance » - Détail de 3 valeurs possibles

La disparité des valeurs représentées dans les corpus d'apprentissage des composants et des modes de défaillance s'explique par la façon dont celui-ci a été construit. En effet, le corpus d'apprentissage a été constitué de manière aléatoire, avec un premier ensemble de 308 FEE puis des FEE analysées au fil de l'eau directement à partir de *PLUS*, car rappelons qu'il ne s'agit pas ici d'automatiser une catégorisation déjà existante dans le processus de la DIPH mais d'introduire une catégorisation nouvelle et de faciliter cela grâce au module de catégorisation de *PLUS*. Néanmoins, l'expert peut désormais s'appuyer sur les autres modules disponibles dans *PLUS* (le moteur de recherche et l'analyse de similarité textuelle notamment) pour choisir les FEE qu'il souhaite ajouter au corpus d'apprentissage pour pouvoir notamment avoir des FEE représentant chacun des composants et des modes de défaillance, et les catégoriser.

5. Conclusion

Ce projet de dépouillement de la base REX SILEX de la DPIH avec l'outil *PLUS*, outil de TAL, a été initié en 2017. L'objectif est d'évaluer la capacité de l'outil *PLUS* à catégoriser les FEE selon les trois champs principaux que sont « Analyse », « Composant » et « Mode de défaillance ». Une première évaluation de l'outil a été réalisée à partir d'un corpus de 200 FEE analysées manuellement par l'expert de la DPIH et les premiers résultats obtenus sont très encourageants. Actuellement, l'outil est capable, avec un taux de réussite moyen de 95%, de distinguer les FEE pertinentes de celles qui ne le sont pas, permettant ainsi d'opérer le premier tri nécessaire au travail d'analyse de l'expert. Rappelons par ailleurs que les erreurs tendent à conserver des FEE dans le périmètre de l'expert plutôt que l'inverse, limitant ainsi le risque d'écarter des FEE nécessitant une analyse approfondie. Pour 86,5% des FEE l'outil est suffisamment sûr de lui pour valider automatiquement les suggestions et le f-score associé à ces valeurs est de 96,3%, soit un taux de réussite tout à fait satisfaisant. Ce premier tri sur le champ « Analyse » permet de filtrer rapidement et de façon fiable les événements significatifs dans le cadre du traitement du REX. Ces FEE font l'objet d'une catégorisation supplémentaire afin d'identifier le composant défaillant parmi les 118 existants et son mode de défaillance parmi les 30 connus. Les résultats sont tels qu'aujourd'hui, près de 9 FEE sur 10 (88%) sont catégorisées correctement pour « Mode de défaillance ». et plus de 3 sur 4 (76%) pour « Composant ».

Aujourd'hui, si les premiers résultats sont satisfaisants, ils permettent également de mettre l'accent sur plusieurs pistes d'amélioration. On observe tout d'abord de façon claire l'importance de la constitution du corpus d'apprentissage ; celui-ci doit être complété et ce de manière à être le plus représentatif possible. Une attention particulière doit notamment être apportée à la présence de FEE pour chacune des valeurs existant dans les référentiels.
Parallèlement, il est nécessaire d'ajuster les seuils permettant que des valeurs soient automatiquement validées afin d'alléger le travail de vérification par l'expert. Pour ce faire, une méthodologie a été définie et a commencé à être mise en place ; elle présente l'avantage de la transversalité : les trois champs étant revus par l'expert, ce sont tous les seuils de validation qui vont pouvoir bénéficier de ce travail et ainsi être ajustés en parallèle, permettant de gagner du temps dans la configuration des modèles et par là même dans l'amélioration des résultats.
Par ailleurs, la bonne qualité générale des résultats encourage à être peut-être moins prudent que ce que la configuration initiale propose. On souhaite notamment alléger la contrainte sur le déclenchement de la catégorisation des champs « Composant » et « Mode de défaillance ». En effet, comme détaillé précédemment 30% de FEE ne sont pas catégorisées dans ces champs car la valeur « Retenue » n'est pas validée automatiquement : l'expert gagnerait à ce que cette contrainte soit supprimée (et/ou le seuil de validation automatique assoupli) afin que des composants et des modes de défaillance soit suggérés.
De plus, bien qu'il ait été peu question d'interface jusqu'à présent, il est à noter que des améliorations sont en cours de développement afin de permettre à l'expert de valider plus facilement et plus rapidement les FEE. Ces ajustements d'ergonomie permettront de réaliser simplement les actions décrites ci-dessus – et plus particulièrement l'amélioration du corpus d'apprentissage – ; l'objectif étant de proposer un outil rapide, fiable et efficace aussi bien dans les suggestions que dans la gestion des modèles de catégorisation.
Notons pour terminer que ce travail d'analyse du REX va permettre d'enrichir la base de données de fiabilité du parc hydraulique de EDF. Il s'agira ainsi, une fois la base *PLUS* complétée de nouvelles FEE, de calculer, pour chaque composant du parc, un taux de défaillance en fonctionnement et/ou une probabilité de défaillance à la sollicitation, données d'entrée des outils de simulation développés à EDF dans le cadre des études de sûreté de fonctionnement.

Références

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J., 2008, « LIBLINEAR: A library for large linear classification. » *Journal of machine learning research*, *9*(Aug), pages 1871-1874.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. « Deep learning. » Nature, 521(7553), pages 436-444.

Martens, D. & Provost, F., 2011, « Explaining Documents' Classifications », in Working paper CeDER. Stern School of Business, New York University.

Porter, M., 2001, « Snowball: A language for stemming algorithms. » http://snowball.tartarus.org/texts/

Raynal, C., Andréani, V., Vasseur, D., Chami, Z. & Hermann, E., 2016, « Apport du Traitement Automatique des Langues pour la catégorisation de Retours d'EXpérience », *Congrès λμ 20*, 11-13 Octobre 2016, Saint Malo, France

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* (Thèse de doctorat), Université de Toulouse II – Le Mirail.