Application du TAL à l'analyse du retour d'expérience à EDF

Coraline Gaucher EDF CIH
Zakarya Chami & Dominique Vasseur EDF R&D

Journée IMdR du 16 mai 2017
Des méthodes aux applications du traitement automatique des langues (TAL) dans le retour d'expérience

SOMMAIRE

CONTEXTE ET OBJECTIF

- **Le traitement du REX collecté sur les parcs de production nécessite des moyens importants car :**
  - Il est volumineux
  - Il doit être validé et complété :
    - Vérifier la cohérence entre champs codés et textes
    - Coder les champs d'intérêt (catégorisation)
- **Intérêt potentiel des outils de TAL pour réduire l'effort nécessaire**
- **Premiers résultats obtenus au cours du projet IMdR P10-5**
  - des progrès réalisé par le TAL depuis les premiers essais réalisés dans les années 90
  - Des résultats intéressants malgré la « mauvaise qualité » des fiches fournies pour les tests
- **Etudes pilotes dans le cadre du programme Innovation EDF R&D**
  - Deux corpus de fiches
    - Hydraulique
    - Nucléaire
  - Test de l'aide à la catégorisation automatique pour montrer l'intérêt de l'approche aux experts REX des directions concernées.

PRÉSENTATION DES ÉTUDES PILOTES ET ENSEIGNEMENTS

REX ANALYSÉS

Données du Nucléaire	Données de l'Hydraulique
Vannes électriques	**Composants d'évacuation des crues dans les barrages hydroélectriques.**
~1 700 fiches d'événement	~2 000 fiches d'événement
2 champs textuels (description des faits)	3 champs textuels (description des faits et des conséquences)
2 métadonnées	8 métadonnées
3 métadonnées à catégoriser : Limite, Mode de Défaillance, Facteur de Dégradation	3 métadonnées à catégoriser : Analyse, Composant, Mode de défaillance
Apprentissage croisé en raison de données d'apprentissage à la fiabilité incertaine	

OUTIL MIS EN ŒUVRE

- **Application PLUS (Processing Language Upgrades Safety)**
  - Web-based application développée
    - pour faciliter l'exploration et l'analyse de bases de données textuelles volumineuses
    - grâce à des techniques de Traitement Automatique des Langues (TAL)
- **Outils disponibles dans PLUS**
  - Module de recherche intelligent
  - Analyse de similarité textuelle (recherche d'antécédents, de signaux faibles)
  - Catégorisation automatique
- **Mesures d'évaluation**
  - Précision = FE automatiquement et correctement attribuées à la valeur v / FE attribuées à la valeur v
    - Met le « bruit » en évidence, les suggestions erronées
  - Rappel = FE automatiquement et correctement attribuées à la valeur v / FE initialement codées v
    - Met le « silence » en évidence, les suggestions manquantes
  - F-score = Combinaison pondérée de la précision et du rappel, permet une bonne évaluation de la fiabilité des suggestions proposées.

ENSEIGNEMENTS

- **Le travail réalisé sur les corpus de fiches fournis a montré l'efficacité de l'aide à la catégorisation.**
  - Amélioration de la qualité des données : Gain de cohérence des données grâce à l'automatisation
  - Gain de temps grâce à la pré-sélection de catégories
    - Experts recentrés sur les cas pertinents
- **Le système n'est pas lié à une taxonomie particulière, il s'adapte à la taxonomie comme au système de reporting.**
  - L'ajout d'une nouvelle valeur est prise automatiquement en charge dès le 1er rapport ainsi catégorisé
  - Changement de catégories maîtrisé : La « reprise des données » peut se faire automatiquement.
- **Nécessité d'un corpus d'apprentissage avec une bonne représentation des valeurs possibles**
- **Aide à prendre du recul sur l'organisation des données**
  - Utilisation de l'outil pour (re)définir la taxonomie

- Très bons résultats si les valeurs sont bien représentées

Analyse - Hydraulique	Suggestion		Total v.ini	Précision	Rappel	F-score
	Non retenue	*Retenue*				
Non retenue	94	6	100	94,7	94,1	**94,4**
Retenue	5	207	212	97,2	97,5	**97,4**
Total Sugg	99	213	312	96,4	96,4	**96,4**

- Forte dégradation des résultats en cas de valeurs très inégalement représentées

Limite – Nucléaire	Suggestion		Total v.ini	Précision	Rappel	F-score
	Hors Limite	*Vanne*				
Hors Limite	3	97	100	50	2,9	**5,6**
Vanne	3	4856	4859	98	99,9	**99**
Total Sugg	6	4953	4959	98	98	**98**

Cas des champs ayant un nombre important de valeurs possibles:

- Champs ‘Type de composant’ : f-score 64,80%
  - 40 valeurs disponibles dont 22 valeurs avec moins de 20 occurrences
  - Qualité de la suggestion d’une valeur dépendante du corpus d’apprentissage : f-score entre 12% et 84%
  - Valeur la plus représentée : ‘Télécommunication’ ; Précision : 74% / Rappel : 95% / f-score : 83%
- Champs ‘Mode de défaillance’ : f-score 57,54 %
  - 29 valeurs disponibles dont 13 valeurs avec moins de 20 occurrences
  - Qualité de la suggestion d’une valeur dépendante du corpus d’apprentissage : f-score entre 9% et 80,5 %
  - Valeur la plus représentée : ‘Absence de signal de télécom’ ; Précision : 68% / Rappel : 98% / f-score : 80,5%

Exemple du mode de défaillance des vannes:

- 4 valeurs organisées en 2 groupes thématiques
  - Les fuites : Fuite externe / Fuite interne
  - Les refus de manoeuvre: Non ouverture / Non fermeture

Résultats :

- Inégaux et devant être améliorés sur les 4 valeurs
- MAIS satisfaisants en considérant les groupes thématiques

➔ **Mise en lumière d'une problématique autour des périmètres des valeurs : sujets à discussion pour les experts et possible réorganisation à l'issue de l'étude**

	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**

	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**

PROJET DE MISE EN ŒUVRE AU CIH

PROJET TAL – ENRICHISSEMENT BASE DE DONNÉES DE FIABILITÉ MATÉRIEL

- **Les objectifs du projet**
  - Dépouiller les fiches (FEE) de la base de REX du parc hydraulique EDF (base SILEX)
    - Dépouillement de la base SILEX brute
    - 1 année ≈ 25 000 FEE
  - Identifier pour chaque FEE, le composant et le mode de défaillance concernés
  - Calculer des données de fiabilité matériel
    - Taux de défaillance en fonctionnement
    - Probabilité de défaillance à la sollicitation
  - Enrichir la base de données de fiabilité matériel
    - Données d'entrée des outils dédiés à la sûreté de fonctionnement des aménagements hydrauliques
    - Evaluation de la fiabilité des composants du parc hydraulique

- **Les attendus de PLUS**
  - Faciliter le travail de dépouillement : gain de temps
  - Améliorer la qualité des données : suggestions à l'analyste, cohérence des données
  - Post-traitement des données : export des données

- **Les phases du projet**
  - Phase 1 : Paramétrage de PLUS avec identification des champs à catégoriser
    - 5 champs à catégoriser
      - Retenue/non retenue
      - Composant
      - Mode de défaillance
      - Poids
      - Durée de réparation
    - 87 composants hydrauliques
    - Modes de défaillances génériques
      - Non réponse à la sollicitation
      - Interruption de fonctionnement
      - Fonctionnement intempestif
      - Pour certains composants : mode de défaillance physique
  - Phase 2 : Phase d'apprentissage
    - Alimentation de PLUS avec un corpus de FEE triées
  - Phase 3 : Phase de validation
    - Test de PLUS sur un volume de FEE brutes, extraites de SILEX
  - .Phase 4 : Dépouillement de l'ensemble de la base SILEX (10 années de REX)
    - Si phase 3 ok
  - A discuter : Phase 5 : Déploiement de PLUS
    - Traitement du REX à venir

- **Projet initié en 2017**
  - Pilotage et financement CIH
- **Projet complexe et ambitieux**
  - Difficulté pour définir la granularité des composants du parc
    - Décomposition fine : facilité d'identification mais statistiques faibles
    - Décomposition macro : perte d'information
  - FEE SILEX
    - Volume important de FEE à traiter (12 années x 25 000 FEE)
    - Beaucoup de champs libre, interprétation parfois délicate
- **Mais projet innovant et riche dont les applications seront, à n'en pas douter, nombreuses !**
  - Etude de sûretés de fonctionnement
  - Analyse de l'accidentologie des vannes
  - Etudes d'optimisation de plans de maintenance

Merci de votre attention