Institut pour la Maîtrise des Risques
Sûreté de Fonctionnement - Management - Cindyniques

Projet P18-2 : " Mise en œuvre de technologies de traitement automatique du langage pour l'interprétation d'un retour d'expérience en langage naturel "

CAHIER DES CHARGES

Version du 22/11/2020

Souscripteurs

CEA	Didier Balestrieri
GRTGaz	Mme Leïla Marle
IRSN	M. Jean-Marie Rousseau et M. Richard Launay
RATP	M. Vincent Ville et M. Vianney Bordeau
SNCF	M. Philippe de Laharpe, Mme Coralie Reutenauer
DGAC	Paul-Emmanuel Thurat, Yoni Malka

Chef de projet

SNCF	M. Philippe de Laharpe

Représentants IMdR

Clément Judek	
André Lannoy	

Table des Matières

1. Présentation et but du projet

1.1. Objet et enjeux

Le projet vise à améliorer la pertinence et l'efficacité du traitement du REX en langage naturel pour aider à la décision et prévenir les erreurs ou incidents.

Plusieurs projets en relation avec le retour d'expérience (REX) et le traitement automatique du langage (TAL) ont été récemment réalisés dans le cadre IMdR. Ces projets ont montré que pour les REX orientés vers la technique et la sûreté de fonctionnement ou pour ceux orientés vers les facteurs organisationnels et humains (FOH), il est nécessaire de travailler sur des données brutes du REX, des données de type texte, soit librement saisies par les opérateurs, soit générées par des systèmes d'automatismes (logs), soit générées par des enregistrements automatiques issus de capteurs (HUMS) pouvant conduire à une situation et donc à un texte libre.

Ces données peuvent être analysées selon plusieurs modalités : par exemple « à froid » pour identifier les causes directes et racines (analyse de fond d'un événement), catégoriser les facteurs contextuels de survenue des événements, classifier les conséquences réelles ou potentielles, ou « à chaud » pour de l'aide à la décision (élaboration d'actions correctives ou préventives, évaluation de leur pertinence, maintenance prévisionnelle par exemple), …

Dans tous les cas la problématique est celle de données nombreuses, non structurées, voire « massives » ; l'analyse du REX sera à mener de plus en plus dans le contexte « *big data* ; nous entendons par *big data* l'ensemble des méthodes de gestion, de traitement et d'interprétation de données « massives » ou fortement hétérogènes, afin de tirer des enseignements valorisables de ces données dans le cadre d'applications métiers.

Ce trop-plein de données peut compromettre la pertinence et l'efficacité de leur traitement « manuel »et nécessite l'utilisation de méthodes statistiques.

Des études récentes (notamment (Tissot, 2017), projet IMdR P10-5) ont montré que la mise en œuvre de ces méthodes nécessite de bonnes connaissances informatiques/méthodologiques/statistiques, de l'expérience pour l'interprétation des résultats, ainsi qu'une réflexion sur les usages métiers et sur les processus organisationnels pour articuler REX et TAL de façon pertinente.

La mise à disposition de tels outils accentue le risque d'effet boîte noire s'il n'est pas accompagné par un guidage à l'interprétation et une montée en compétences des analystes (experts).

En complément des résultats déjà acquis lors des études précédentes, l'objectif premier de cette étude est de partager et de s'approprier les conditions de mise en œuvre de méthodes de Traitement Automatique du Langage (notamment de méthodes statistiques) pour le traitement et l'interprétation d'un retour d'expérience en langage naturel.

1.2. Résultats attendus

Le résultat principal du projet est un guide de recommandations pratiques concernant l'utilisation des méthodes statistiques pour le TAL.

Ces recommandations porteront notamment sur les dimensions :

- Stratégiques : typologie d'objectifs, d'usages et de technologies à mobiliser pour être en phase avec les besoins et attentes de différents profils d'utilisateurs
- Techniques : caractéristiques des données d'entrée (volumétrie, format, gestion documentaire et niveau de dématérialisation, homogénéité des données, besoins de ressources lexico-sémantiques complémentaires), niveau de modélisation préalable des connaissances métiers sur les événements et la description des incidents, formats de restitution visuelle des résultats (datavisualisation)
- Organisationnelles et culturelles : compétences requises, accompagnement du changement pour mettre en œuvre de nouvelles pratiques liées au traitement et à l'analyse du REX
- Financières : coût pour le test, le déploiement et la maintenance de solutions TAL ; évaluation et mode de calcul du retour sur investissement / estimation de la valeur apportée par les différentes solutions TAL

Ces recommandations permettront d'identifier les prérequis pour l'implémentation de solutions de type TAL dans les entreprises et les facteurs de succès pour la mise en œuvre de telles solutions.

Elles seront étayées par des exemples de cas réels traités et interprétés concernant les retours d'expérience technique, humain et organisationnel, fournis par les souscripteurs ou par des sources libres d'accès. L'étude et le guide feront l'objet d'un rapport de synthèse, d'applications réelles illustratives, d'un résumé de synthèse et d'un jeu de diapositives.

1.3. Tâches proposées

Ce programme est donné à titre indicatif. Il sera explicité dans une proposition qui intégrera les différents besoins des souscripteurs exprimés dans le présent cahier des charges, ainsi que les domaines d'applications souhaités**.**

Tâche 0 : Lancement du projet – Rappel des besoins des souscripteurs – Objectifs de l'expérimentation et choix de jeux de données tests.

Les jeux de données candidats comporteront au moins un jeu de données publiquement accessibles à tous les souscripteurs. Ils pourront être complétés par des données de retour d'expérience spécifiques aux souscripteurs.

Ces données permettront d'approfondir différentes fonctionnalités liées au TAL des retours d'expérience, telles que :

- Le traitement de données hétérogènes de retour d'expérience, on peut à coup sûr avoir besoin de données structurées ou quantitatives,
- La classification et la catégorisation automatique selon un « modèle » préétabli ou par inférences inductives (*machine learning*),
- La discrimination statistique de signaux (repérage d'événements rares ou nouveaux, de signaux faibles, ou à l'inverse, de récurrences),
- La détection de tendances et d'inflexions,
- L'analyse de similarités et de corrélations entre plusieurs événements,
- La recherche d'informations textuelles,
- La comparaison à des documents prescripteurs,
- Etc.

Tâche 1 : Recueil d'informations sur les méthodes de traitement candidates, état de l'art bibliographique.

Ces méthodes candidates peuvent être (Tanguy, 2017 ; Tissot, 2017) :

- Les méthodes de sémantique latente (LSA : *Latent Semantic Analysis* ; LSI : *Latent Semantic Indexation*),
- Le *topic modeling*,
- L'analyse distributionnelle (*word embeddings*, plongements lexicaux),
- D'éventuelles autres méthodes, comme par exemple le gradient *boosting tree*, peuvent être adaptées à différentes fonctions recherchées, ou d'autres méthodes plus « anciennes » comme celle de l'analyse de données,

avec prise en compte des méthodes nouvelles des trois dernières années, qu'elles soient stabilisées ou en cours d'émergence.

Cet état bibliographique devra préciser : les concepts de la méthode, les hypothèses, les caractéristiques des données traitées, les outils logiciels disponibles (commerciaux ou open source), les applications existantes et leurs résultats, les avantages et inconvénients relatés dans la documentation technique, l'intelligibilité des résultats …

Tâche 2 : Recueil d'informations sur les prétraitements de données

Cette tâche adressera les questions posées par la préparation (prétraitement) des données source : Faut-il un prétraitement ? Faut-il des ressources (lexique, terminologie, liste d'acronymes, règles génériques…) ? Que peuvent apporter des prétraitements sémantiques ? Le prétraitement enrichit –il ou au contraire perturbe-t-il l'interprétation ?...

A noter que le prétraitement est garant de la qualité des données et donc de la qualité des résultats. Il est donc essentiel dans une démarche d'utilisation d'outils statistiques pour le traitement du TAL.

Tâche 3 : Traitement des jeux de données choisis par les souscripteurs.

Dans cette tâche on effectuera un prétraitement du (des) jeu(x) de données choisi(s) parmi des jeux de données proposés. On spécifiera les difficultés rencontrées. On spécifiera aussi les éventuels avantages, apports et limites de ces prétraitements.

Puis on effectuera le traitement proprement dit. Une attention particulière sera portée sur la présentation des résultats dans le but de faciliter leur lecture et leur interprétation. Des recommandations seront alors énoncées.

Tâche 4 : Interprétation des résultats obtenus – Conclusions quant à l'apport des outils TAL.

Cette interprétation sera effectuée par le prestataire avec l'appui des experts des souscripteurs. Les représentations en sortie des outils TAL présentent-elles des difficultés d'interprétation pour les experts du REX ? Peut-on retourner aux données brutes initiales ou intermédiaires afin de juger la pertinence des résultats ? Comment la connaissance des experts (métiers, statisticiens, fiabilistes, …) intervient-elle dans cette analyse ?

Cette tâche couvrira aussi la question de l'accessibilité des méthodes aux différents profils d'utilisateurs des données du retour d'expérience. Les utilisateurs visés sont : l'expert du retour d'expérience, l'opérationnel, le décideur… On examinera en particulier : l'utilisation des outils logiciels, l'interprétation, la confiance dans les résultats obtenus…

Tâche 5 : Conclusions, synthèse et perspectives de R&D

La synthèse se positionnera sur les quatre dimensions (stratégique, technique, organisationnelle, financière) mentionnées à la section 1.2 du présent CdC.

1.4. Quelques références

- Journée IMdR (16 mai 2017), Des méthodes aux applications du TAL dans le REX.
- Tanguy Ludovic (2017), Quel TAL pour le retour d'expérience ? Réflexions sur les besoins et les solutions actuelles.
- Tissot Claire (16 mai 2017), *Text mining* sur des données d'accidentologie.
- Mason L., Baxter J., Bartlett P.L., Frean M. (1999), *Boosting Algorithms as Gradient Descent, Advances in Neural Information Processing Systems*, MIT Press, pp512-518.
- Projet IMdR (2013) – Détection de signaux faibles
- Projet IMdR (2013) – Retour d'expérience et TAL (Traitement Automatique des Langues)
- Projet IMdR (2017) – HUMS (*Health and Usage Monitoring Systems*)
- Blatter C., Raynal C. (2014), *Méthodes d'analyse textuelle pour l'interprétation des rex humais, organisationnels et techniques,* congrès λμ 19, Dijon, octobre 2014
- Christian Blatter, Pascal Tonnerre, Adeline Pernet, Stéphanie Donnet, Coralie Reutenauer, *et al*. *Vers un REX ferroviaire prenant en compte les avancées en matière de facteurs humains et organisationnels*, Congrès λμ 21 " Maîtrise des risques et transformation numérique : opportunités et menaces ", Oct. 2018, Reims, France. ffhal-02075359f
- Christian Blatter, Pascal Tonnerre, Stéphanie Donnet, Coralie Reutenauer, Cécile Million-Rousseau. *Traitements linguistiques pour la recherche d'information et l'analyse en FOH de REX ferroviaires*.

Congrès λμ 21, " Maîtrise des risques et transformation numérique : opportunités et menaces ", Institut pour la Maîtrise des Risques, Oct. 2018, Reims, France. ffhal-02063576f

2. Organisation du projet

2.1. Méthode de travail, réunions d'avancement et durée du projet

Les études et réflexions sont menées par le prestataire du contrat qui prend en compte, dans toute la mesure du possible, les points de vue des membres du groupe de projet. Il est tenu d'argumenter sa position en cas de divergences et de non prise en compte des observations.

L'état d'avancement des études est présenté au groupe de projet lors des réunions d'avancement. 6 réunions de projet sont à prévoir, soit :

- une réunion de lancement,
- une réunion d'avancement à l'issue de la tâche 1,
- une réunion d'avancement à l'issue de la tâche 2,
- une réunion d'avancement à l'issue de la tâche 3,
- une réunion d'avancement à l'issue de la tâche 4,
- une réunion finale (à l'issue de la tâche 5).

La première réunion (réunion de lancement) du groupe de souscripteurs est mise à profit pour expliciter le cahier des charges du projet, préciser la méthode de travail et établir le calendrier des réunions futures.

Une réunion d'avancement peut englober la présentation de plusieurs tâches.

La réunion finale est aussi très importante. Les résultats et les documents associés y sont présentés, validés et acceptés par les souscripteurs.

La durée de réalisation du projet est de douze mois (12 mois) à compter de la date de lancement du projet.

2.2. Composition du groupe de projet

Le groupe de projet comprend :

- deux animateurs, le chef de projet et le représentant du prestataire,
- les représentants des souscripteurs, membres de droit,
- des membres de la société prestataire,
- les représentants de l'IMdR,
- Eventuellement, des experts du domaine, avec l'accord des membres du groupe de projet.

2.3. Documents produits au cours du projet

Chaque réunion fait l'objet d'un compte rendu succinct rédigé par le prestataire, hormis celui de la réunion de lancement qui sera rédigé par le chef de projet. Ce compte rendu rassemble tous les renseignements relatifs au fonctionnement du groupe et à l'organisation des réunions : participants à la réunion et thèmes évoqués, lieu, date, ordre du jour et date de la réunion suivante.

Les résultats du projet comprennent :

- les rapports intermédiaires à l'issue de chaque tâche,
- le rapport final,
- les modèles utilisés, les données des cas d'étude et leurs résultats,
- d'éventuelles macros logicielles produites lors du traitement,
- un résumé de synthèse de trois pages,
- les comptes rendus de réunion,
- un jeu de diapositives au format « power point » de présentation du rapport final.

2.4. Confidentialité

Les membres de la société qui réalisent le travail et ceux du groupe des souscripteurs sont tenus de respecter la règle de confidentialité des démarches, des méthodes, des données et des résultats. Ils sont tenus de respecter à la lettre les clauses de confidentialité figurant dans les contrats de partenariats signés à l'issue de la réunion de lancement du projet. Une charte de mise en œuvre des projets IMdR reprenant les règles de confidentialité sera annexée au contrat. Elle est consultable sur le site de l'IMdR.

Ainsi, les accès aux bases de données de REX et aux comptes rendus d'événements sécurité, les publications des résultats d'analyse sont soumises à l'autorisation de chacune des sociétés partenaires.

Les comptes rendus et les rapports ne pourront donc être diffusés hors du groupe des souscripteurs sans leur autorisation expresse ainsi que celle de l'IMdR.

Cas d'application soumis au prestataire

Il est précisé que le prestataire s'engage à une obligation de stricte confidentialité couvrant toutes les informations qu'il sera amené à connaître dans le cadre des discussions et présentations.

Pour les besoins du présent article, l'expression "Information Confidentielle" signifie toute information, quelle que soit la forme sous laquelle elle se présente (orale, écrite, magnétique, électronique, graphique ou numérique), contenant, montrant ou consistant en une information ou une documentation de nature technique qui ne doit pas être divulguée.

Le prestataire s'oblige vis-à-vis des souscripteurs à une obligation stricte et générale de confidentialité en ce qui concerne toute Information Confidentielle communiquée dans le cadre du projet.

Il est expressément convenu entre les parties que les Informations Confidentielles et leurs reproductions, éventuellement transmises par les souscripteurs au prestataire, restent la propriété des souscripteurs et que les communications d'Informations Confidentielles faites en vertu du projet ne pourront en aucun cas être interprétées comme conférant de manière implicite, au prestataire, une quelconque licence portant sur les droits de propriété intellectuelle ou industrielle des souscripteurs en relation avec les Informations Confidentielles, que ces droits existent au jour de la conclusion du contrat ou qu'ils naissent ultérieurement.

Le prestataire s'engage à ne faire aucune reproduction et à ne divulguer à aucun tiers tout ou partie des Informations Confidentielles qui lui auront été communiquées.

Ces informations sont accessibles aux seules personnes qui, en raison de leur compétence et/ou de leur fonction, participent au traitement des cas d'application soumis.

2.5. Lieu des réunions

Les réunions se dérouleront, selon les choix émis par les membres du groupe projet, dans les locaux de l'IMdR, dans les locaux du prestataire, dans les locaux de l'un des souscripteurs, ou par visioconférence.

2.6. Rôle du prestataire

Le rôle du prestataire consiste à :

- réaliser les tâches du projet, conformément au présent Cahier des Charges,
- participer aux réunions et à l'animation du groupe de projet,
- donner son avis sur tous les sujets abordés, relevant de ses compétences et faire part des difficultés rencontrées,
- rédiger les comptes rendus des réunions d'avancement,
- rédiger les rapports intermédiaires associés aux différentes tâches du projet,
- faire un jeu de diapositives présentant les résultats du projet,
- rédiger le rapport final,

- rédiger le résumé de synthèse (trois pages au plus) du projet.

Les documents seront fournis au format Word ou PowerPoint pour permettre au chef de projet de suivre facilement les modifications et aux souscripteurs d'apporter facilement leurs commentaires.

2.7 Rôle du chef de projet

Le chef de projet veille à la finalisation du cahier des charges et à son acceptation par les souscripteurs. Il convoque la première réunion, dite de lancement, et en rédige le compte rendu.

Par ailleurs, il prend en charge l'organisation des réunions du groupe de projet :

- définition des ordres du jour et des convocations aux réunions en accord avec le prestataire,
- invitation des membres occasionnels, sur demande du groupe,
- approbation des comptes rendus de réunion et des rapports de projet.

En plus de son rôle d'animateur, le chef de projet veille au bon déroulement du projet, en s'assurant que les diverses tâches prévues, techniques et administratives, sont bien remplies et, par conséquent, que le but du projet est atteint dans les délais impartis.

3. Contenu attendu des offres de prestation

3.1. Description et organisation des activités

L'offre de prestation devra décrire l'ensemble des activités prévues pour obtenir les résultats attendus du projet.

L'offre de prestation précisera les moyens humains et matériels à mettre en œuvre pour obtenir la bonne réalisation des activités, ainsi que l'expertise et les références dans le domaine. Concernant les moyens humains, une attention particulière devra être apportée quant à la définition des niveaux de compétences auxquels le prestataire s'engage à faire appel ainsi que la disponibilité effective de ces moyens pour tenir les engagements de délais.

3.2. Documents livrables

Les documents listés au paragraphe 2.3 sont considérés comme les documents livrables contractuellement par le prestataire.

3.3. Jalonnement des travaux

Le prestataire doit présenter dans sa proposition un calendrier des activités qu'il envisage de conduire en utilisant les orientations proposées aux chapitres 1 et 2 du présent cahier des charges.

3.4. Critères de qualité des offres

Les offres des différents candidats à la sous-traitance pour ce projet seront évaluées en fonction des critères de choix habituels de l'IMdR :

- la compréhension critique du projet proposé,
- la pertinence technique de l'offre,
- les références et compétences du(des) proposant(s) dans le domaine de l'analyse de risques, des évaluations probabilistes, des incertitudes aussi bien d'un point de vue industriel que méthodologique,
- les délais et le montant de l'offre.

Le (s) CV de la (des) personne(s) impliquée(s) sera (seront) fourni(s).

3.5. Conditions financières

Les propositions financières devront détailler, pour chacune des tâches, la nature des intervenants, les volumes des interventions et les coûts annexes prévus. Elles devront aussi préciser l'échéancier des paiements attendus.