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

Méthodes d'analyse textuelle pour l'interprétation des REX humains, organisationnels et techniques

---

Synthèse du rapport final

Projet de l'IMdR n° P10-5

Chef de Projet :
Monsieur Christian BLATTER, SNCF

Contractant :
CFH – Conseil en Facteurs Humains

Conseil en Facteurs Humains

IMdR – 12 avenue Raspail – 94250 GENTILLY
Tél. 33 (0)1 45 36 42 10 Fax. 33 (0)1 45 36 42 14
www.imdr.fr - contact@imdr.eu

L'**Institut pour la Maîtrise des Risques** tient à remercier :

- **Monsieur Christian BLATTER, SNCF** qui a dirigé cette étude,
- **les sociétés** qui ont souscrit à ce projet **et leurs collaborateurs** qui ont participé à sa réalisation :
  - EADS — **M. Richard LEBLOND**
  - EDF — **Mme Dominique VASSEUR**, **Mme Aurélie LEGER**
  - GDF Suez — **M. Youness HSSAINI**, **Mme Leïla MARLE**
  - INERIS — **M. Damien FABRE**
  - RATP — **M. Christian VIZENEUX**
  - SNCF — **M. Christian BLATTER**
- son **Délégué Technique, M. John OBAMA** et son **Vice-Président M. André LANNOY** qui ont contribué cette étude.

L'**Institut pour la Maîtrise des Risques** tient également à remercier les sociétés suivantes, sociétés auditionnées dont les outils de TAL ont subi des tests sur des corpus de données de retour d'expérience des industriels :

- ONTOTEXT — **M. Atanas KIRYAKOV**
- ONTOLOGOS — **Mme Cécile MILLION-ROUSSEAU**

- PEPITE **M. Philippe MACK**
- RAPID-I **M. Sebastian LAND**
- TEMIS **M. Fabien GAUTHIER**

L'**Institut pour la Maîtrise des Risques** tient enfin à remercier les responsables techniques qui ont mené le projet à bien Vanessa ANDREANI et Céline RAYNAL (CFH), Cécile FABRE et Ludovic TANGUY (Université Toulouse-Le Mirail, Laboratoire CLLE-ERSS, UMR 5263).

Note de présentation et de synthèse

Le projet IMdR P10-5, intitulé « Méthodes d'analyse textuelle pour l'interprétation des REX humains, organisationnels et techniques », a impliqué pendant un an les souscripteurs EADS, EDF, GDF Suez, l'INERIS, la RATP et SNCF. Il a pour objectif de définir de nouvelles approches, fondées sur le Traitement Automatique des Langues (TAL), pour analyser et interpréter les documents textuels relatifs au retour d'expérience (REX).

Ce projet a été scindé en quatre phases distinctes ; nous les détaillons ci-après :

1. Dans un premier temps, un état de l'art a été constitué, à la fois sur le REX en regard du TAL et sur le TAL vis-à-vis du REX.

2. Puis un benchmark a été mené afin de sélectionner des outils de TAL potentiellement pertinents dans l'analyse de REX.

3. Cette première sélection a été suivie d'une évaluation pratique des outils sélectionnés à partir de données réelles, appartenant à des corpus constitués par les souscripteurs.

4. Enfin, des préconisations ont été émises afin de permettre aux souscripteurs, et plus largement aux acteurs du REX, de faire des choix raisonnés quant aux traitements de TAL qu'ils souhaitent intégrer à leur processus de REX.

L'état de l'art porte sur des aspects ciblés du REX et du TAL ainsi que sur leurs interactions. Ainsi, le retour d'expérience est considéré comme un processus de signalisation, de stockage et d'analyse des événements (écarts, défaillances, incidents) dans une entreprise, mais également comme un objet textuel qui va, par là-même, être objet du TAL. En parallèle, le TAL est présenté en tant qu'ensemble de techniques adaptables aux problématiques du REX. De manière synthétique, on peut dire que du point de vue du TAL, le REX va être pris en compte comme un objet fait de texte qui s'intègre dans un processus de traitement relativement large, et qui, en résumé, consiste en trois grandes étapes :

- l'alimentation de la base de données de REX,
- la sélection de données spécifiques au sein de cette base de données et leur validation avant analyse,
- l'analyse des données sélectionnées.

Une liste de techniques et de méthodes de TAL a donc été établie, en fonction de leur pertinence vis-à-vis des différentes phases de traitements pouvant être appliqués aux données de REX, et donc des besoins des acteurs du REX. Par exemple, la catégorisation automatique de rapports de REX peut se révéler adéquate si l'on souhaite organiser une base de données en fonction des types d'événements ayant eu lieu dans l'entreprise. Dans ce cas précis, le TAL intervient donc dès l'intégration des événements dans la base de données[1], afin d'attribuer une catégorie à chacun

[1] On peut également envisager de faire intervenir la catégorisation automatique sur des rapports déjà présents dans la base mais n'ayant jusque-là pas encore fait l'objet d'une première analyse (nécessaire pour établir un lien entre l'événement et une catégorie).

d'entre eux. D'autres techniques permettent de faciliter la tâche des analystes à d'autres moments du cycle de vie du REX : pour la sélection de documents en vue d'une analyse par exemple. L'ensemble de ces techniques sont décrites et illustrées dans l'état de l'art ; nous les explicitons rapidement ici :

- la **catégorisation automatique** permet de déduire automatiquement une ou des catégories à partir d'un contenu textuel ;

- la **vérification de la cohérence** permet de valider l'adéquation entre un contenu narratif et une catégorie, un champ contrôlé associé, mais également de comparer des documents relatifs à un même événement et de valider leur adéquation ;

- la **recherche d'information** permet de structurer le contenu textuel des REX composants une base de données, d'interroger celle-ci grâce à différents types de requête, et de renvoyer une liste ordonnée de rapports correspondants ;

- le **calcul de similarité** aide à identifier des rapports similaires à un rapport de référence dans une base de données prise comme objet d'étude ;

- la **fouille de texte** (ou Text Mining) se base sur une formalisation des textes qui permet de les interroger et de faire émerger des informations nouvelles dans les contenus textuels et d'identifie des corrélations entre champs contrôlés et textes narratifs ;

- le **clustering** permet de faire émerger des groupes de documents sur la base de critères communs et peut faire apparaître une organisation implicite de ceux-ci.

D'autres fonctionnalités sont transversales et permettent de dépasser les limites d'une analyse des seuls mots pour prendre en compte la richesse et le sens des expressions :

- l'application de modèles linguistiques met en évidence des **relations sémantiques spécifiques** (telles que la causalité, la temporalité, la subjectivité du locuteur, etc.) ;

- la **constitution de ressources langagières** a pour vocation de structurer la connaissance terminologique et ontologique d'un domaine donné ; les ressources ainsi créées seront un puissant moyen d'améliorer les traitements futurs sur les données textuelles.

Les fonctionnalités correspondant le plus aux besoins des souscripteurs ont été sélectionnées pour la suite de l'étude. L'enjeu le plus important est généralement de tirer l'information pertinente d'une masse de documents qu'il est difficile d'appréhender manuellement.

Après avoir listé les types d'applications pertinentes pour traiter le REX, la deuxième phase du projet a consisté à répertorier une quinzaine d'outils proposant les fonctionnalités prioritairement attendues par les souscripteurs, qu'ils soient issus de la recherche académique ou de l'industrie. L'expression du besoin des souscripteurs a permis de classer les objectifs de traitement par ordre de priorité et deux des fonctionnalités ont été unanimement jugées non prioritaires : l'aide à la rédaction et la vérification de cohérence inter-documents (i.e. l'agrégation de document). La liste des outils a été présentée aux souscripteurs, et une concertation commune a permis de sélectionner six outils qui ont ensuite fait l'objet d'une analyse plus poussée (nous y revenons ci-après). Les critères majeurs qui ont été pris en compte étaient :

- l'adéquation des prestations proposées avec les besoins exprimés par les souscripteurs et les apports concrets escomptés ;
- la validité des techniques de TAL employées ;

- la facilité de mise en place de ces outils pour une éventuelle installation en contexte industriel pour l'analyse du REX ;
- la facilité de manipulation pour les futurs utilisateurs, non spécialistes du TAL ni de l'informatique.

La troisième étape a donc permis de tester pratiquement les six outils sélectionnés. Tous les souscripteurs ayant des données de REX ont constitué un corpus issu de leur propre base de données de défaillances ou d'incidents, caractéristiques du retour d'expérience industriel. De cette façon, les spécificités de chacun des types de données ont été prises en compte. Les outils et logiciels testés étaient de deux types : d'une part, les outils voués à une seule fonctionnalité (des logiciels dédiés uniquement à la catégorisation automatique ont par exemple été testés), et d'autre part, les outils multifonctions, qui s'apparentent à des plateformes généralement assez puissantes et qui couvraient plusieurs besoins d'un même souscripteur. Dans tous les cas, il a été fait appel aux éditeurs de ces outils pour réaliser les tests, puisqu'aucun ne fournissait de démonstrateur directement utilisable en ligne. Pour cela, un court cahier des charges a été rédigé pour chacun des souscripteurs(les demandes n'étant pas forcément les mêmes d'un souscripteur à l'autre) à destination d'un ou plusieurs des éditeurs impliqués. Chaque éditeur a donc été chargé de réaliser des tests sur les données des souscripteurs, et de fournir une prestation ciblée sur les demandes de chacun. Les résultats de ces tests ont fait l'objet d'une restitution réalisée par chacun des éditeurs, et, dans certains cas, d'une version de démonstration manipulable. Ces évaluations ont permis de constater l'adaptabilité des techniques de TAL aux problématiques spécifiques des souscripteurs du projet, mais également de tirer des constats d'ordre général concernant l'interaction entre TAL et REX :

- Les techniques de TAL ont effectivement leur place dans le traitement et l'analyse du REX : dans la plupart des cas, les outils testés sont non seulement pertinents par rapport à la demande initiale, mais également efficaces et performants dans le traitement des données. Ce résultat est d'autant plus encourageant que les textes de REX ont à ce jour peu fait l'objet d'analyse textuelle automatique.

- Ces techniques nécessitent dans la plupart des cas un investissement en temps dans la mise en place d'un système les utilisant : il n'existe pas (ou peu), à notre connaissance, d'outils de TAL sur étagère, c'est-à-dire utilisables immédiatement. C'est d'ailleurs pour cette raison qu'il est très rare de trouver des démonstrateurs en ligne dans lesquels injecter des données de test pour évaluation. Cependant, le coût engendré par la mise en place de tels outils est compensé par le gain de temps et de performance une fois les logiciels installés, paramétrés et entraînés.

- Il est important d'avoir une idée relativement précise de la demande, afin d'appliquer les traitements adéquats et de ne pas « se perdre » dans les différentes fonctionnalités.

- Chaque configuration besoin/outil est particulière et implique un certain nombre de contraintes, notamment en termes de données disponibles et de contexte ; il est important de les prendre en compte dans l'intégration et l'utilisation d'un outil de TAL.

Suite à cette évaluation, des préconisations ont été émises lors de la quatrième et dernière étape de l'étude. Elles sont destinées aux souscripteurs du projet, mais ont également une portée plus large et peuvent s'étendre à l'ensemble de la communauté du REX. En effet, en fonction des enjeux, des besoins et des attentes de chaque acteur impliqué dans le REX, mais aussi des contraintes organisationnelles et techniques où il se place, les implications seront différentes et les outils ne

seront pas toujours les mêmes, ou ne seront pas utilisés de la même manière. C'est pourquoi les recommandations prennent en compte différents critères.

Le premier d'entre eux s'exprime en termes de contraintes pesant sur les données de REX destinées à être exploitées par les traitements de TAL envisagés. D'un point de vue technique comme d'un point de vue linguistique, les documents doivent être adaptés à un traitement informatique. Par exemple, les documents doivent se présenter dans un format exploitable par la machine, et être en nombre suffisant. De même, plus les données seront « propres », plus leur exploitation sera efficace. Le second critère est celui de la disponibilité des experts : dans tous les cas, un certain laps de temps devra être alloué à la mise en place des outils choisis, et un ou des experts du domaine devront être disponibles pour participer à leur déploiement. Enfin, la nature des données va influencer le choix des traitements à effectuer : si les documents sont constitués à la fois de texte libre et de valeurs prédéfinies, comme des catégories par exemple, les traitements envisagés pourront être différents de ceux choisis dans le cas où les données sont constituées uniquement de texte libre.

Chaque technique de TAL identifiée comme pertinente lors de l'étude est donc abordée dans les recommandations en fonction de ces différents critères : les contraintes spécifiques à un module de traitement sont répertoriées. Les avantages à utiliser la technique en question sont également recensés, de façon à ce que chaque acteur du REX puisse faire des choix en toute connaissance de cause.

Cette étude a permis d'explorer les différentes façons dont le TAL peut faciliter et enrichir l'analyse du retour d'expérience. L'état de l'art a mis en lumière les techniques et méthodes potentiellement pertinentes pour aider à cette analyse. Les deux phases de benchmark des outils ont permis de déterminer une liste d'outils efficaces sur les données de REX, mais également de montrer qu'en pratique, le TAL apporte effectivement une plus-value. D'autre part, le classement de ces outils et logiciels au sein d'une typologie permettra à l'avenir d'y intégrer d'autres outils potentiellement intéressants, et de savoir dans quel contexte ils peuvent être utilisés. Enfin, les recommandations recensent l'ensemble des critères permettant d'opérer un choix raisonné dans le cadre de la sélection d'outils de TAL adaptés à une problématique donnée, avec ses contraintes et les gains escomptés en termes d'efficacité. De manière générale, le TAL peut donc apporter des solutions efficaces et concrètes au traitement du REX d'autant plus performantes que le retour d'expérience est important en taille et nécessite donc de lourds traitements et analyses. Les outils faisant appel à ces techniques ne peuvent être envisagés que dans un contexte donné, avec des contraintes clairement identifiées et des besoins bien définis. Une fois le contexte déterminé, c'est l'ensemble des acteurs du REX qui peut tirer profit des apports du TAL et des analyses qu'il permet, dans le but de mieux comprendre et évaluer les risques humains, organisationnels et techniques et ainsi de mieux réduire leur impact.