Risque et TAL : détection, prévention, gestion
Introduction au 1er atelier

Natalia Grabar[1] Ludovic Tanguy[2]
(1) STL - Savoirs Textes Langage : CNRS et Université de Lille
UMR 8163, Lille, France
(2) CLLE-ERSS - Equipe de Recherche en Syntaxe et Sémantique : CNRS et Université de Toulouse
UMR 5263, Toulouse, France
natalia.grabar@univ-lille3.fr, tanguy@univ-tlse2.fr

RÉSUMÉ

Nous présentons ici le premier atelier *Risque et TAL* portant sur les méthodes de traitement automatiques des langues pour la détection, la prévention et la gestion des risques. Les travaux présentés dans le cadre de cet atelier sont issus de travaux académiques mais aussi d'applications développées par des acteurs industriels. Ils couvrent les principaux domaines pour lesquels la notion de risque est au centre de préoccupations de par l'ampleur des conséquences à éviter : biomédical (médecine et pharmacologie), chimie et transports, mais abordent aussi des aspects plus transversaux de l'activité humaine, comme les environnements professionnels et les spécifications. Ces différents travaux montrent à la fois la diversité des données visées (retours d'expérience, réseaux sociaux, publications scientifiques, enquêtes, documentation technique), les objectifs des analyses (extraire de l'information liée aux risques, contrôler ou vérifier les ambiguïtés) et les solutions techniques (recueil de données, analyse de corpus, développement de ressources).

ABSTRACT

Risk and NLP: detection, prevention, management. Introduction to the first workshop.

This article is the introduction to the first workshop dedicated to *Risk and NLP*, addressing the use of natural language processing methods for the detection, prevention and management of risk. The papers presented during the workshop come from both academic and industrial actors. They cover the most risk-prone domain such as biomedicine (medicine and pharmacology), chemistry and transportation, but also address more transversal issues of human activity such as professionnal environments and technical documentation and requirements. The works presented also show the variety of the processed data (intervention reports, social network communications, academic papers, surveys, technical documentation), the objectives of the analyses (extraction of information related to the risk, ambiguity control, documentation checking), and of technical solutions (data collection, corpus analysis, resources development).

MOTS-CLÉS : Risque, méthodes de TAL.

KEYWORDS: Risk, NLP methods.

1 Contexte

La prise de risques est propre au fonctionnement des personnes et de la société. Ceci est entre autre nécessaire pour faire repousser les frontières de la connaissance actuelle et pour effectuer des progrès et des avancées scientifiques et industriels, bien que d'autres contextes, plus néfastes pour les humains et la société, peuvent aussi être concernés. Le risque touche une grande variété de situations et de domaines. De nombreuses activités peuvent ainsi rencontrer des situations anormales ou inattendues dont les conséquences peuvent être fortement dommageables. Parmi les domaines d'activités concernés, nous pouvons mentionner ceux liés à:

— transport (aérien, routier, ferroviaire, maritime, spatial),
— énergie (centrales électriques, industrie pétrolière),
— biomédical (médecine, pharmacologie, pharmacovigilance),
— industrie lourde (métallurgie, chimie, construction).

De tels domaines sont plus particulièrement exposés de par l'ampleur des dommages possibles. De ce fait, les acteurs de ces domaines ont mis en place des systèmes de vigilance afin de détecter, prévenir et limiter les risques. La plupart de ces systèmes font intervenir une dimension langagière importante (rapports d'accidents ou d'incidents, retours d'expérience des interventions, veille sur la base des flux de communication, etc.), si bien que le TAL peut y contribuer à la gestion du risque (Tulechki, 2011; Tulechki & Tanguy, 2012). Cependant, dans de nombreux autres domaines, la détection de situations à risque peut prendre appui sur l'analyse de données textuelles ou de communications souvent massives et/ou spécifiques. On voit notamment que cela peut concerner les réseaux sociaux (Neubig *et al.*, 2011; Collier, 2011) ou la littérature scientifique (Blake, 2004; Hamon *et al.*, 2010; Blanchemanche *et al.*, 2013; Grabar & Kerry, 2015).

Depuis quelques années, on voit apparaître un ensemble de travaux et d'applications du TAL proposant des solutions ou des pistes de réflexions dans ce sens. Notons par exemple :

— la journée d'étude "*Linguistique et Traitement Automatique des Langues pour l'Aéronautique et l'Espace : Dimensions langagières du risque*" (Toulouse, 2014) [1]
— le projet P10-5 "*Méthodes d'analyse textuelle pour l'interprétation des REX humains, organisationnels et techniques*" commandité par l'institut pour la maîtrise des risques (IMdR) [2](Blatter & Raynal, 2014)
— l'atelier "*Computational Methods in Pharmacovigilance*" [3] (Pise, 2012).

Ces travaux sont généralement menés et diffusés au sein de communautés thématiques correspondant aux domaines d'application. L'objectif de cet atelier est donc de proposer un espace d'échange transeversal entre différents acteurs travaillant sur la détection, la description, l'analyse ou la prévention du risque et utilisant des méthodes et ressources de TAL pour aider la réalisation de tâches visées.

Parmi les approches considérées on citera de façon non exhaustive :

— analyse des retours d'expérience (rapports d'accident/incident) : assistance aux experts ;
— détection des risques potentiels dans les rapports d'activité : identification des "signaux faibles" ;
— détection des risques potentiels dans les communications : identification du risque explicite dans les communications entre usagers d'un système et/ou experts d'un domaine ;
— aide à la communication ou la rédaction de documents critiques : définition et vérification de langages contrôlés ou de consignes de rédaction ;

1. http://w3.erss.univ-tlse2.fr/textes/seminaires/lingTALaerospace.html
2. http://www.imdr.eu/
3. http://natalia.grabar.perso.sfr.fr/CMPV/

— développement de ressources pour la prévention/gestion des risques : extraction d'information et ingénierie des connaissances métier.

La volonté de cet atelier est de réunir une communauté large à la fois en termes de domaines et d'applications, dans l'idée d'identifier quels sont les besoins véritables des acteurs du terrain et les réponses que les techniques actuelles du TAL peuvent fournir. De ce fait, cet atelier s'adresse autant aux chercheurs qu'aux acteurs associatifs et industriels. Cette ouverture se traduit par l'implication dans le comité de programme de personnalités non académiques mais impliquées dans la gestion du risque dans différentes structures (institutions professionnelles et entreprises).

2 Comité de programme

Nous remercions vivement les membres du comité de programme de l'atelier dont la participation a permis d'effectuer l'expertise des soumissions reçues et d'en améliorer la qualité.

3 Déroulement de l'atelier

Sur les neuf soumissions reçues pour l'atelier *Risque et TAL*, huit ont été acceptées: six comme présentations longues et deux comme présentations courtes.

Nous voulons remercier tous les auteurs qui ont soumis leurs travaux à l'atelier.

L'atelier s'est déroulé en suivant le planning ci-dessous :

— 14h00-14h15 : *Introduction à l'atelier*, Natalia Grabar et Ludovic Tanguy
— 14h15-14h30 : *Méthodes d'analyse textuelle pour l'interprétation des REX humains, organisationnels et techniques*, Celine Raynal et Christian Blatter
— 14h30-14h55 : *Vers la définition de nouvelles langues contrôlées*, Anne Condamines
— 14h55-15h20 : *Information extraction from the social media: a linguistically motivated approach*, Nelleke Oostdijk, Ali Hürriyetoğlu, Marco Puts, Piet Daas and Antal van Den Bosch
— 15h20-15h45 : *Exploitation de différentes approches pour détecter et catégoriser le risque chimique et bactériologique*, Natalia Grabar et Thierry Hamon
— 15h45-16h00 : *Prévention des risques liés à l'environnement de travail : constitution d'un corpus oral en vue de son traitement automatique*, Sandra Cestic et Iris Eshkol-Taravella
— 16h00-16h30 : pause café
— 16h30-16h55 : *Types de risque médical et leur traitement avec des méthodes de TAL*, Natalia Grabar et Frantz Thiessard
— 16h55-17h20 : *Identification de termes flous et génériques dans la documentation technique : expérimentation avec l'analyse distributionnelle automatique*, Émilie Merdy, Juyeon Kang et Ludovic Tanguy
— 17h20-17h45 : *PLUS : pour l'exploration de bases données REX*, Celine Raynal, Assaf Urieli et Eric Hermann
— 17h45-18h15 : Discussion finale

4 Contenu des communications

Les soumissions acceptées et présentées lors de l'atelier montrent la variabilité et la richesse des différents aspects liés au risque. Elles indiquent également un grand potentiel d'ouverture et de progression dans les travaux futurs de la communauté.

Le travail présenté par Christian Blatter et Céline Raynal *Méthodes d'analyse textuelle pour l'interprétation des REX humains, organisationnels et techniques* concerne le traitement des rapports de REX (retour d'expérience des incidents), qui comportent une quantité croissante de descriptions textuelles d'événements de sécurité ou de défaillances techniques. Ainsi, un projet IMdR P10-5 a permis d'identifier des méthodes et outils de Traitement Automatique des Langues permettant d'exploiter rapidement une très grande quantité de rapports constitués de verbatim : catégorisation automatique, identification d'événements similaires, recherche d'information, etc. Il s'agit donc d'un travail assez générique, dont le bénéfice peut être ressenti dans différents domaines liés au risque qui doivent faire face à des volumes importants de données textuelles.

L'intervention *Vers la définition de nouvelles langues contrôlées* d'Anne Condamines repose sur plusieurs études réalisées depuis une dizaine d'années à CLLE-ERSS sur les langues contrôlées (CNLs), préconisées pour la rédaction de la documentation technique afin d'en assurer une certaine normalisation. Le principal constat de ce travail est que les CNLs ne sont pas toujours adaptées pour un type de documents techniques donné et facilement utilisables. Par ailleurs, l'impact réel de leur mise en œuvre sur l'amélioration de la lisibilité de ces documents a été très peu évalué. L'article dresse un panorama des problèmes associés aux CNLs et propose de nouvelles pistes de constitution des CNLs. Dans cet objectif, certaines méthodes de TAL et de psycholinguistique pourraient être mises en oeuvre pour améliorer les CNLs existantes ou en proposer de nouvelles. Là aussi, il s'agit d'une réflexion de fond dont peuvent profiter de nombreux domaines et tâches afin d'assurer une meilleure prévention et gestion du risque.

Le travail *Information extraction from the social media: a linguistically motivated approach* de Nelleke Oostdijk, Ali Hürriyetoğlu, Marco Puts, Piet Daas et Antal van Den Bosch propose une méthode flexible pour l'extraction de l'information sur le trafic routier à partir des réseaux sociaux. L'abondance de micropostes sur Twitter rend en effet possible d'identifier ce qui se passe sur les routes puisque les utilisateurs y rapportent en temps réel ce qu'ils sont en train d'observer. Cette information est très pertinente et peut aider les organisateurs de la sécurité routière et les conducteurs à être mieux préparés et à prendre les décisions appropriées. Les auteurs distinguent 22 catégories d'information supposées être pertinentes dans le domaine du trafic routier. Une précision de 74 % est obtenue avec les tweets individuels ; les auteurs considèrent cette performance satisfaisante, d'autant plus qu'il existe habituellement plusieurs tweets sur un événement donné, ce qui offre une meilleure possibilité de détecter l'information pertinente. Ce travail montre entre autre les avantages liés à l'exploitation de l'information contenue dans les réseaux sociaux, de même qu'au fait d'y trouver des informations dupliquées et mises à jour constamment.

Le travail *Exploitation de différentes approches pour détecter et catégoriser le risque chimique et bactériologique* présenté par Natalia Grabar et Thierry Hamon relate la détection des informations qui concernent les risques liés aux substances chimiques et bactériologiques avec différentes méthodes de TAL. L'objectif consiste donc à proposer une aide automatique pour l'analyse de la littérature scientifique afin de détecter les phrases indicatives du risque que présentent les substances chimiques ou des bactéries. La tâche est abordée comme un problème de catégorisation : il s'agit de catégoriser les phrases des textes dans les classes du risque lié aux substances. Trois approches sont utilisées

et évaluées : à base de règles, apprentissage supervisé et recherche d'information. De meilleurs résultats sont obtenus avec l'apprentissage supervisé et la recherche d'information. En fonction des approches, les résultats obtenus montrent jusqu'à 0,8 de F-mesure. Il s'agit d'un travail qui propose une comparaison entre trois méthodes de TAL et deux domaines liés au risque, et qui peut également fournir des indices à d'autres domaines qui font face à de grands volumes de données textuelles.

La communication *Prévention des risques liés à l'environnement de travail : constitution d'un corpus oral en vue de son traitement automatique* de Sandra Cestic et Iris Eshkol-Taravella présente la constitution d'un corpus oral destiné à l'étude de l'expression verbale de la perception de facteurs physiques dans les environnements de travail. Cette étude a pour objectif d'apporter les connaissances nécessaires à la finalisation du développement d'une application informatique dédiée à la prévention des nuisances physiques au travail générées par le bruit et les ambiances thermiques. Les auteurs abordent la méthodologie mise en œuvre pour collecter des données orales authentiques et constituer un corpus susceptible d'anticiper au mieux les procédures de traitement automatique pour l'extraction d'informations relatives aux risques. Il s'agit d'un travail qui vise la prévention du risque abordée alors d'une manière originale grâce à la constitution et exploitation d'un corpus authentique collecté auprès des salariés, en particulier en ce qui concerne la perception de facteurs physiques (bruit et température) dans un environnement de travail. Ce travail peut donner des indices sur la qualité de vie et de travail des salariés, et par conséquence d'améliorer leurs conditions de travail.

De manière plus ciblée, le travail *Types de risque médical et leur traitement avec des méthodes de TAL* de Natalia Grabar et Frantz Thiessard aborde plus précisemment le risque médical. D'une part, une classification de risques médicaux est proposée. Le domaine médical est en effet intimement lié avec le risque : la médecine a pour vocation de traiter les patients qui présentent un risque ou une suspicion d'une pathologie donnée et par ailleurs, le processus de soins médicaux peut également générer un risque pour les patients. D'autre part, une analyse plus détaillée de certains types de risques est proposée, en particulier ceux qui sont traités avec des méthodes de TAL: les facteurs de risque et les risques liés aux traitements médicaux et au séjour à l'hôpital.

L'étude *Identification de termes flous et génériques dans la documentation technique : expérimentation avec l'analyse distributionnelle automatique* de Émilie Merdy, Juyeon Kang et Ludovic Tanguy se place dans le cadre du développement des ressources linguistiques utilisées par un système de vérification automatique de documentations techniques. L'objectif est d'étendre semi-automatiquement des classes de termes intrinsèquement flous ainsi que des termes génériques afin d'améliorer le système de détection de passages ambigus reconnus comme des facteurs de risque. Les auteurs mesurent et comparent l'efficacité de méthodes d'analyse distributionnelle automatiques en comparant les résultats obtenus sur des corpus de taille et de degré de spécialisation variables pour une liste réduite de termes d'amorce. Ils montrent en particulier que si un corpus de taille trop réduite est inutilisable, son extension automatique par des documents similaires donne des résultats complémentaires à ceux que produit l'analyse distributionnelle sur de gros corpus génériques. Il s'agit d'un travail dont l'utilité est avérée lors de la préparation des méthodes et outils pour le traitement automatique des textes liés au risque.

Finalement, un dernier travail *PLUS : pour l'exploration de bases données REX* de Celine Raynal, Assaf Urieli et Eric Hermann propose la plateforme PLUS (*Processing Language Upgrades Safety*), qui exploite différents modules basés sur des techniques de Traitement Automatique des Langues, afin d'effectuer une recherche "intelligente", un calcul de similarité textuelle et une catégorisation automatique des rapports du retour d'expérience, tels que ceux-ci sont recueillis à grande échelle dans des industries liées au transport, à l'énergie ou au domaine médical. Là encore, les méthodes de

TAL permettent de traiter de grands volumes de données textuelles, tout en assurant leur traitement homogène et systématique. L'objectif des outils présentés ici est d'assister le travail d'analyse des experts, qui restent au centre de tout dispositif de gestion des risques à large échelle, et pour lesquels les apports du TAL sont désormais cruciaux face à la masse de données textuelles.

Références

BLAKE C. (2004). A text mining approach to enable detection of candidate risk factors. In *Medinfo*, p. 1528–1528.

BLANCHEMANCHE S., RONA-TAS A., DUROY A. & MARTIN C. (2013). Empirical ontology of scientific uncertainty: Expression of uncertainty in food risk analysis. In *Society for Social Studies of Science*, p. 1–27.

BLATTER C. & RAYNAL C. (2014). Méthodes d'analyse textuelle pour l'interprétation des rex humains, organisationnels et techniques. In *Congrès Lambda Mu 19*, Dijon, France.

COLLIER N. (2011). Towards cross-lingual alerting for bursty epidemic events. *J Biomed Semantics*, **2**(5), S10.

GRABAR N. & KERRY N. (2015). Deux approches pour catégoriser le risque. In *EGC 2015, RNTI-E-28*, p. 83–88, Luxembourg.

HAMON T., GRAÑA M., RAGGIO V., GRABAR N. & NAYA H. (2010). Identification of relations between risk factors and their pathologies or health conditions by mining scientific literature. In *Medinfo*, p. 964–8.

NEUBIG G., MATSUBAYASHI Y., HAGIWARA M. & MURAKAMI K. (2011). Safety information mining: What can NLP do in a disaster ? In *5th International Joint Conference on NLP (IJCNLP)*, p. 965–973, Chiang Mai, Thailand.

TULECHKI N. (2011). Des outils de tal en support aux experts de sûreté industrielle pour l'exploitation de bases de données de retour d'expérience. In *RECITAL*.

TULECHKI N. & TANGUY L. (2012). Effacement de dimensions de similarité textuelle pour l'exploration de collections de rapports d'incidents aéronautiques. In *TALN*, p. 439–446.