Page de titre

Vous pouvez indiquer ici le titre du livre et le nom des auteurs.

Titre ouvrage :

Détection du risque de substances chimiques dans les documents textuels

Sous titre (éventuel) :

Nom de(s) auteur(s) :
Niña Kerry, Natalia Grabar

Nous reprendrons cette page pour sa mise en page.

Page des mentions légales

Nous remplissons cette page.

Table des matières

Remerciements

Nous remercions Sandrine Blanchemanche et Laura Maxim pour les données annotées que nous utilisons dans notre travail.

Le travail présenté ici est effectué dans le cadre du projet PNRPE DicoRisque financé par le Ministère de l'écologie, que nous voudrions également remercier.

Partie 1

Détection du risque de substances chimiques dans les documents textuels

Le risque chimique couvre les situations où les produits chimiques sont ou peuvent être dangereux pour la santé humaine, animale et pour l'environnement. Si certains produits et substances sont maintenant clairement identifiés comme dangereux *(e.g.*, arsenic, listeria), nos connaissances actuelles sur d'autres substances sont moins complètes. Dans de telles situations, il est possible que différents points de vue existent sur une question donnée, ce qui peut mener à l'apparition de situations de controverse. Par exemple, une substance comme le bisphénol A entre dans la composition de différentes matières, y compris les plastiques d'emballage et les bouteilles plastiques pour les boissons. Cette substance est donc très présente dans les produits de consommation courante. Le risque lié au bisphénol A, qui est un perturbateur endocrinien (c'est-à-dire qu'il affecte le système hormonal, comme le font également les phtalates), varie entre autre selon la durée d'exposition, la dose, la masse corporelle et l'âge des sujets. Toutes ces substances ont un impact grave pour la santé. La détection des informations qui concernent le risque des substances chimiques occupe donc une place importante auprès des agences environnementales et les chercheurs qui y consacrent leur activité. Cependant, d'une part la profusion de données et d'autre part les controverses qui les concernent créent une situation où il devient difficile de trouver rapidement et efficacement les informations pertinentes.

Dans notre travail, nous proposons une méthode automatique pour le traitement de gros volumes de données textuelles et pour la détection des phrases qui expriment les dangers et les incertitudes liés aux substances chimiques. Nous travaillons avec des classifications qualitatives qui décrivent les facteurs de l'incertitude liés au risque chimique et alimentaire. Les méthodes exploitées proviennent de la recherche d'information : nous voulons faire le rapprochement entre les classes du risque (considérées comme les requêtes) et les phrases

qui se trouvent dans les documents traités (considérées comme les réponses à ces requêtes). Nous traitons la littérature scientifique et institutionnelle car c'est la source typique des connaissances actuelles sur les substances chimiques et leur impact sur les organismes et l'environnement. Ce type de documents est exploité par les organismes réglementaires, qui sont en charge des autorisations de mise sur le marché de différents produits et de la marchandise. Malgré les incertitudes connues sur les substances et les produits, ces organismes doivent néanmoins être en mesure de prendre des décisions quant à leur commercialisation (Wardekker et al., 2008 ; van der Sluijs et al., 2008). Par ailleurs, nous utilisons aussi des données de référence, créées manuellement par les experts.

Chapitre 1 : Données et approche

La méthode, qui repose sur des outils de la recherche d'information, exploite les particularités sémantiques des écrits scientifiques. Cela consiste en combinaison des informations scientifiques factuelles, qui concernent les faits scientifiques d'expériences et des observations effectuées suite à ces expériences, et des information sur les incertitudes et les imprécisions, qui peuvent exister autour de ces expériences et interprétations (Hyland, 1995 ; Mauranen, 1997 ; Hamm, 1991 ; Witterman et al., 2003 ; Dhami et al., 2005).

1.1 Corpus

Les corpus traités proviennent de la littérature scientifique et sont représentatifs des données typiques utilisées pour la prise de décisions. Ainsi, pour l'étude du risque alimentaire, nous exploitons 115 rapports institutionnels dédiés à plusieurs substances dangereuses à la consommation (e.g., arsenic, BSE, TSE, dioxine, listeria, ammeline, mélamine, nitrates, salmonelle). Les parties pertinentes (résumé, introduction, conclusion) de ces rapports contiennent plus de 240 000 mots. Pour l'étude du risque chimique, nous utilisons un rapport de l'EFSA sur le bisphénol A. Ce rapport contient plus de 80 000 mots.

1.2 Classifications

Nous utilisons deux classifications des facteurs du risque (chimique et alimentaire). Ces classifications décrivent différents aspects potentiellement liés à et révélateurs de la nocivité des substances et de leur danger. Chaque classe reçoit un libellé et une définition. La classification du risque alimentaire (Blanchemanche et al., 2011) propose des classes comme *epistemic uncertainty, inference animal human, data, model, missing factors/variables, surrogate data, measurement, causal inference, arbitrary assumptions of model, surrogate population, surrogate context, sampling, surrogate hazard agent, measure*. La classification du risque chimique (Maxim et al., 2014) est inspirée de la précédente et couvre également plusieurs aspects (e.g., *form of the dose-effect relationship, choice of the causal mechanisms for interpreting the findings, sensitivity of the essay, choice of the dose tested, natural/unexplained variability, sample size, performance of measurement, control of confounders, reproducibility*). Voilà quelques exemples de phrases relatives au risque et typiques de celles que l'on trouve dans les écrits scientifiques : *The authors reported a positive correlation with age, although the the authors themselves underlined that the small sample size (11 women) strongly limited the value of this association (sample size) ; The Panel is not aware of any clearly reproducible adverse effect expressed specifically at low BPA doses only (reproducibility) ; Modelling*

assumptions and inter-individual variability within all dose groups were accounted for by applying a calculated CSAF (arbitrary assumptions model) ; Because of a lack of toxicity data in domestic animals, EFSA provisionally recommends to apply this tolerable intake level as established for humans also to domestic animals (inference animal human).

1.3 Données de référence

Les données de référence sont obtenues suite aux annotations manuelles des rapports par les spécialistes en évaluation du risque. Les annotations du risque alimentaire sont effectuées dans le cadre du projet Met@risk (*https://www6.jouy.inra.fr/metarisk*), alors que les annotations du risque liée au bisphenol A sont effectuées dans le cadre du projet DicoRisque. Les annotations du risque alimentaire sont disponibles en ligne sur le site du projet. Nous avons 1 836 et 425 phrases annotées respectivement dans les deux corpus.

1.4 Ressources linguistiques

Les ressources linguistiques sont utilisées pour enrichir les requêtes (libellés des classes) et pour collecter ainsi plus de réponses (phrases provenant des rapports institutionnels). Ces ressources contiennent donc des synonymes ou des termes et mots équivalents. Nous exploitons des ressources existantes (UMLS, 2010) et des ressources construites de manière non supervisée (Brown, 1992 ; Liang, 2005) à partir des corpus traités. Dans ce dernier cas, nous utilisons des méthodes distributionnelles (Harris, 1968, Brown, 1992), qui permettent de grouper les mots qui partagent des contextes similaires au sein de clusters. Il est supposé en effet que les mots qui partagent de tels contextes partagent également la sémantique commune ou proche, ce qui est calculé avec des fréquences, l'information mutuelle et les mesures de similarité (e.g., Jaccard ou Cosine) (Curran, 2004). La finesse des clusters dépend de leur nombre : nous faisons plusieurs expériences en générant entre 200 et 600 clusters sur les deux corpus traités.

1.5 Approche pour la détection du risque dans les documents textuels

Le système de recherche d'information Indri (Strohman et al., 2005) est au cœur de notre approche. Ce système est basé sur le modèle statistique de la langue (e.g., champ aléatoire de Markov, graphes, n-grammes). Indri fournit plusieurs fonctionnalités que nous exploitons, comme par exemple :

- les raciniseurs Porter (Porter, 1980) et Krovetz (Krovetz, 1993), qui permettent d'enlever les finales de mots jugées comme non pertinentes (pluriels, féminins, désinences des formes fléchies de verbes, etc.) ;
- la combinaison de mots clés (les mots clés de la requête avec éventuellement leurs synonymes ou mots équivalents) ;
- la pondération des mots clés (avec normalisation par la taille des documents, avec tfidf, okapi, etc.)

Les réponses renvoyées par Indri sont classées dans l'ordre de leur pertinence par rapport à la requête. Il est ainsi possible de ne retenir que 10, 20 ou 30 premières réponses. Ceci peut en effet rendre les résultats plus facilement analysables par un analyste humain.

Lors des expériences, le corpus de travail entier de même que les annotations sont partagées en deux ensembles : ensemble d'entraînement et ensemble de test. Les paramètres du système sont ainsi réglés avec l'ensemble d'entraînement. De cette manière l'évaluation est effectuée sur un ensemble indépendant. Lors de l'entraînement, nous réglons les fonctionnalités de Indri (utilisation des raciniseurs, réglage des poids de mots, combinaison des mots, influence des ressources linguistiques), et nous effectuons également la sélection de mots clés.

Dans notre travail, les mots clés de la requête pour une classe donnée sont issus du libellé de cette classe. Par exemple, pour la classe *sample size* les mots-clés sont *sample* et *size*. En cas d'utilisation de ressources linguistiques, les mots qui se trouvent dans les mêmes clusters que les mots clés sont ajoutés à la requête. Ces mots sont vérifiés pour leur spécificité (de savoir s'ils génèrent suffisamment de précision) et sensibilité (de savoir s'ils génèrent suffisamment de couverture dans les résultats). Lorsque le mot n'est pas spécifique à la classe donnée et lorsque la spécificité n'est pas suffisante, ce mot n'est pas retenu pour cette classe. L'utilisation de ressources linguistiques permet d'augmenter la sensibilité des résultats pour une classe donnée.

Les résultats sont évalués avec les mesures de précision (spécificité), rappel (sensibilité), F-mesure (moyenne harmonique de précision et de rappel), MAP (mean average precision) calculée avec N premiers résultats retenus (10, 20, 30, etc).

Notre baseline consiste à effectuer la recherche de phrases pertinentes pour une classe donnée sans effectuer des traitements supplémentaires (pondération, racinisation, ressources linguistiques, etc.). Seuls les mots de la classe sont pris en compte.

Chapitre 2 : Expériences et résultats obtenus

Au total, 24 classes relatives au risque alimentaire et 23 classes relatives au risque chimique sont traitées. Pour certaines classes, nous ne trouvons pas de phrases pertinentes, ce qui est dû au fait que peu de phrases sont annotées pour une classe donnée et l'évaluation ne peut pas être faite correctement (pas d'exemples dans l'ensemble de test) ou bien au fait que le contenu de la classe n'est pas suffisamment spécifique et ne peut pas être distingué correctement. Pour les classes qui peuvent être traitées, les performances varient entre 0.1 et 1 (les meilleurs résultats possibles).

Dans le tableau 1, nous présentons les résultats globaux obtenus sur les deux corpus traités. Nous pouvons voir qu'en moyenne les résultats sont meilleurs avec le corpus du risque chimique. Il est possible que les libellés des classes sont plus explicites dans la classification du risque chimique.

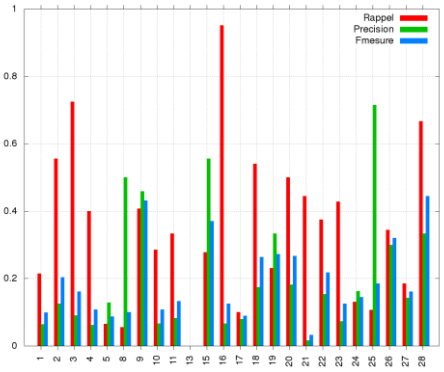
	R. alimentaire	*R. chimique*
baseline	0,18	0,20
Krovetz okapi	0,199	0,219
Krovetz tfidf	0,199	0,219
Porter okapi	0,191	0,20
Porter tfidf	0,191	0,20
Krovetz clusters	0,226	0,32

TAB 1 – ***Performances de F-mesure obtenues sur les deux corpus traités : différents paramétrages, valeurs moyennes.***

Dans la suite, nous présentons les résultats au niveau des classes individuelles.

2.1 Détection automatique du risque alimentaire

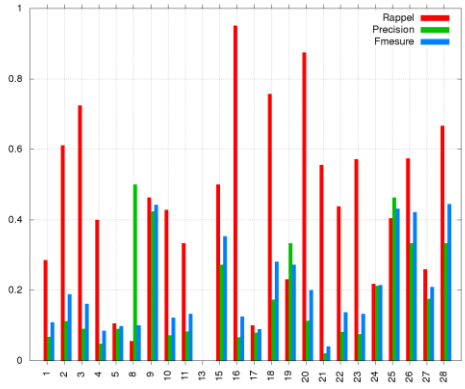
A la figure 1, nous présentons les performances en termes de précision (en vert), rappel (en rouge) et F-mesure (en bleu) obtenues avec la baseline. Nous pouvons observer que les résultats sont assez variables selon les classes et que certaines classes reçoivent peu de réponses. Dans les expériences suivantes, nous allons voir l'influence de la racinisation et de la pondération des mots clés.



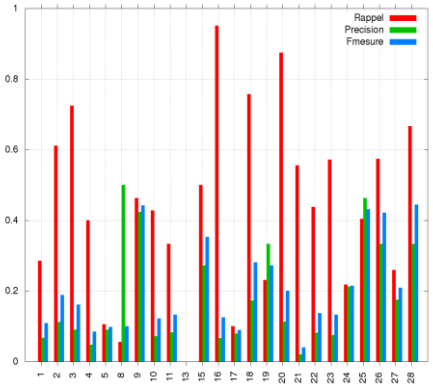
Baseline : paramètres par défaut

Fig. 1 — Performances (précision rappel et F-measure) obtenues lors de la détection automatique du risque alimentaire. Pas de paramètres spécifiques (baseline).

Comme attendu, avec l'utilisation des raciniseurs de Krovetz (figure 2) et de Porter (figure 3), ce sont surtout les valeurs du rappel qui augmentent. Les valeurs de la précision peuvent diminuer pour certaines classes. En revanche, les performances globales sont améliorées de plusieurs points. L'utilisation de la pondération des mots clés (tfidf, okapi) ne modifie pas la F-mesure, mais améliore les valeurs de la MAP. En effet, les phrases retournées sont les mêmes, mais leur ordre change et devient plus correct.

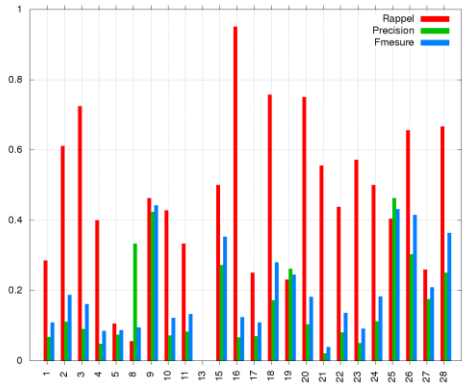


Krovetz okapi

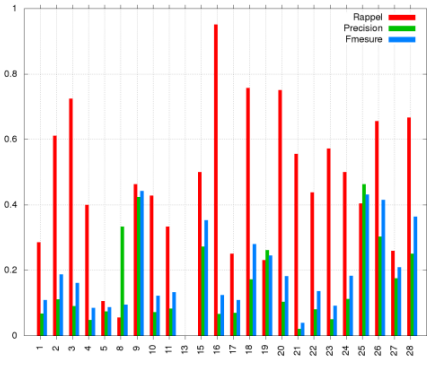


Krovetz tfidf

Fig. 2 — Performances (précision rappel et F-measure) obtenues lors de la détection automatique du risque alimentaire. Paramètre : racinisation avec Krovetz, pondération avec okapi et tfidf



Porter okapi

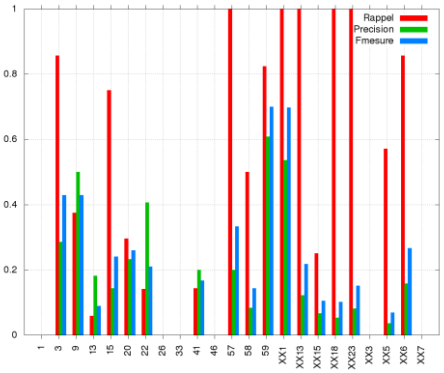


Porter tfidf

Fig. 3 – Performances (précision rappel et F-measure) obtenues lors de la détection automatique du risque alimentaire. Paramètre : racinisation avec Porter, pondération avec okapi et tfidf

2.2 Détection automatique du risque chimique

A la figure 4, nous présentons les performances en termes de précision (en vert), rappel (en rouge) et F-mesure (en bleu) obtenues avec la baseline. Nous pouvons observer que les résultats sont meilleurs que ceux obtenus avec le risque alimentaire. Nous pensons que les libellés des classes sont plus explicites et spécifiques et permettent ainsi que collecter plus de réponses correctes.

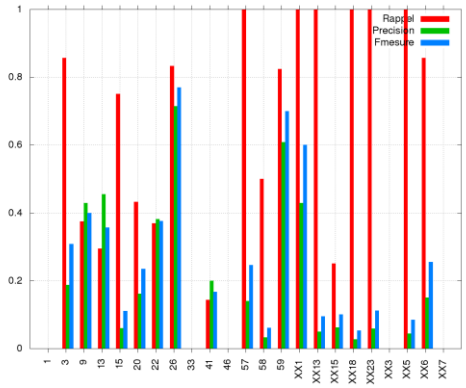


Baseline : paramètres par défaut

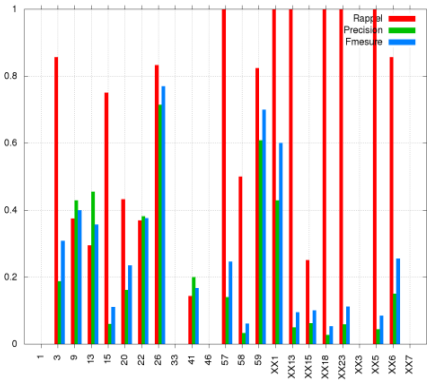
Fig. 4 – Performances (précision rappel et F-measure) obtenues lors de la détection automatique du risque chimique BPA. Pas de paramètres spécifiques (baseline).

Comme auparavant, avec l'utilisation des raciniseurs de Krovetz (figure 5) et de Porter (figure 6), ce sont surtout les valeurs du rappel qui augmentent. Les valeurs de la précision

peuvent diminuer pour certaines classes. En revanche, les performances globales sont également améliorées de plusieurs points. Nous observons la même évolution de la MAP.

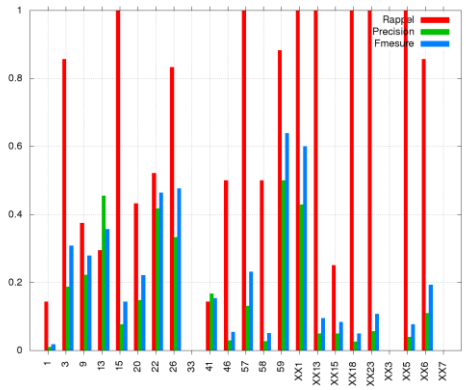


Krovetz okapi

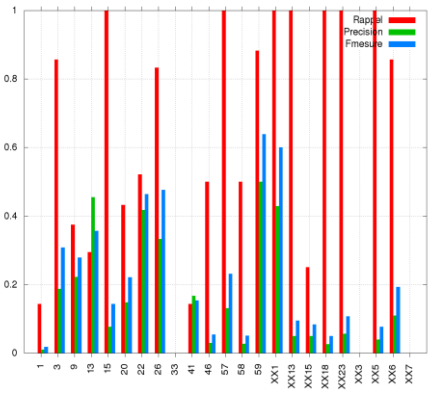


Krovetz tfidf

Fig. 5 – Performances (précision rappel et F-measure) obtenues lors de la détection automatique du risque chimique BPA. Paramètre : racinisation avec Krovetz,pondération avec okapi et tfidf



Porter okapi



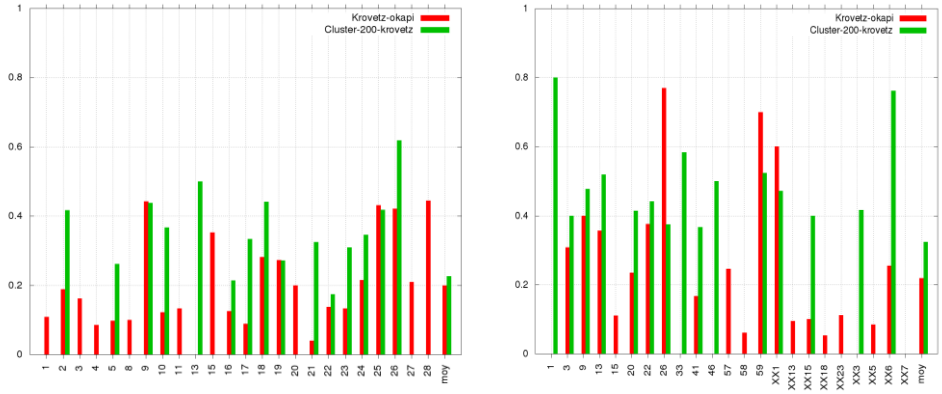
Porter tfidf

Fig. 6 – Performances (précision rappel et F-measure) obtenues lors de la détection automatique du risque chimique BPA. Paramètre : racinisation avec Porter,pondération avec okapi et tfidf

2.3 Influence de ressources linguistiques

A la figure 7, nous montrons l'influence de l'utilisation de ressources linguistiques : les résultats en vert sont obtenus avec les ressources linguistiques (clusters de mots) et le raciniseur Krovetz, en rouge seule l'utilisation du raciniseur est effectuée. Nous pouvons voir que l'impact varie en fonction des classes : il est positif pour certaines mais négatif pour d'autres. Globalement, le rappel augmente mais les performances globales sont

légèrement détériorées. Nous pensons que les ressources linguistiques sont à utiliser pour augmenter la couverture des réponses du système. Par contre, lorsque la spécificité est recherchée, il est recommandé de ne pas utiliser de ressources supplémentaires.



R. alimentaire *R. chimique*

Fig. 7 – Variation de performances (MAP) en fonction de l'utilisation ou non de ressources linguistiques : Krovertz-okapi en rouge, Krovetz-cluster en vert.

2.1 Bilan des expériences

Nous avons présenté quelques expériences autour du traitement automatique des rapports institutionnels et la détection de phrases relatives au risque induit par les substances chimiques. La méthode proposée repose sur un système existant de recherche d'information dont nous testons plusieurs fonctionnalités. Nous utilisons aussi des pré-traitements spécifiques et utilisons des ressources linguistiques générées de manière non-supervisée. Deux classifications du risque sont utilisées : une décrivant le risque alimentaire et une autre le risque chimique. Les classes de ces classifications décrivent différents aspects pouvant être impliqués dans la détection du risque pendant les expériences chimiques et biologiques, comme par exemple les méthodes de mesure, la contamination, les données traitées et leur taille, la variabilité des données, la représentativité de l'échantillon, etc. Dans notre approche, les libellés des classes sont considérés comme les requêtes alors que les phrases des documents sont considérées comme les réponses à ces requêtes. Plusieurs rapports sont traités avec l'approche proposée.

Les résultats montrent que nous obtenons de meilleures performances avec la classification du risque chimique testée sur le corpus BPA et que les performances sont moins bonnes avec les rapports consacrés au risque alimentaire. L'influence de différents paramètres (raciniseurs, pondération des mots clés, ressources linguistiques) sont variables selon les classes. De manière générale, ces paramètres permettent d'augmenter le rappel, mais diminuent la précision. En fonction des corpus, les performances globales (F-mesure) augmentent (avec les raciniseurs) ou diminuent (avec les ressources linguistiques). Nous pensons que ces paramètres sont à varier selon que l'on cherche à augmenter la couverture ou la spécificité des résultats. La pondération permet d'améliorer les valeurs de la MAP, ce qui est également un point positif de nos expériences.

Nous pouvons faire une comparaison avec une approche basée sur l'apprentissage supervisé. Avec l'apprentissage supervisé, la taille de données doit être plus importante pour avoir des résultats intéressants. Par exemple, dans un travail précédant (Grabar et al., 2014), nous avons pu obtenir une F-mesure atteignant jusqu'à 0.92 pour certaines classes du risque chimique. Cependant, de telles performances ne sont possibles qu'avec des classes qui ont un nombre d'exemples suffisamment élevé. Dans l'expérience indiquée, seulement 8 classes ont pu être traitées. Des résultats comparables sont également obtenus avec le corpus du risque alimentaire (Blanchemanche et al., 2011). Par comparaison à ces expériences, une approche basée sur la recherche d'information est assez indépendante du volume de données de référence disponible : il est possible de régler le système sur un plus faible volume de données. Par exemple, nous avons traité 24 classes relatives au risque alimentaire et 23 classes relatives au risque chimique, ce qui dépasse largement la couverture des expériences en apprentissage supervisé.

Il est aussi apparu que les données de référence, par rapport auxquelles est effectuée l'évaluation, ne sont pas complètes. Avec notre méthode, nous pouvons trouver d'autres phrases pertinentes mais qui ne figurent pas dans les données de référence. Actuellement, nous complétons cet aspect. Le travail avec des données de référence plus complètes peut permettre d'améliorer les performances globales du système. Parmi d'autres perspectives, nous voulons aussi exploiter d'autres ressources linguistiques et tester d'autres combinaisons de mots clés. Les résultats générés peuvent être évalués par les experts en risque chimique et vue de juger de leur utilité pour l'analyse des rapports sur le risque chimique et pour la prise de décisions.

Bibliographie

BLANCHEMANCHE S., BUCHE P., DIBIE-BARTHELEMY J., FAINBLATT MELEZE E., IBANESCU L., RONA-TAS A. Ontology building: an application in food risk analysis. Proc of TIA 2009.

BLANCHEMANCHE S., RONA-TAS A., CORNUÉJOLS A., DUROY A., MARTIN C. An Ontology of Scientific Uncertainty: Methodological Lessons from Analyzing Expressions of Uncertainty in Food Risk Assessment. Tech. report 2011

BROWN PF., de SOUZA PV., MERCER RL., Della PIETRA VJ., LAI JC. Class-Based n-gram Models of Natural Language. *Computational Linguistics* 1992, **18**(4) :467-479.

CURRAN JR. From distributional to semantic similarity. University of Edinburgh. 2004, PhD thesis.

DHAMI MK., WALLSTEN TS. Interpersonal comparison of subjective probabilities: towards translating linguistic probabilities. *Memory & Cognition* 2005 **33**(6):1057-1068.

GRABAR N., WANDJI TCHAMI O., MAXIM L. Machine learning-based detection of chemical risk. Proc. of MIE 2014 : 725-729.

HAMM RM. Selection of verbal probabilities: a solution for some problems of verbal probability expressions. *Organizational behavior and human decision processes*. 1991 **48**: 193-223.

HARRIS ZS. *Mathematical Structures of Language*. Wiley, 1968. New York, NY, USA

HYLAND K. The Author in the Text: Hedging in Scientific Writing. *Hong Kong papers in linguistics and language teaching*. 1995 **18**: 33-42.

KROVERTZ R. Viewing Morphology as an Inference Process. Proc. of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1993 :191-202.

LIANG P. Semi-Supervised Learning for Natural Language. Massachusetts Institute of Technology, Boston, USA. 2005. Master thesis.

MAURANEN A. Hedging in Language Revisers' Hands. Hedging and discourse: approaches to the analysis of a pragmatic phenomenon in academic texts. 1997. Walter de Gruyter: 115-133.

MAXIM L., Van Der SLUIJS JP. Qualichem In Vivo: A Tool for Assessing the Quality of In Vivo Studies and Its Application for Bisphenol A. *PLOS one*. 2014.

PORTER M. An algorithm for suffix stripping. *Program* 1980, **14**(3) :130-137.

STROHMANT., METZLER D., TURTLE H., CROFT WB. Indri: a language-model based search engine for complex queries. Tech. rep. Proc. of the International Conference on Intelligent Analysis, 2005.

UMLS. Knowledge Sources Manual. National Library of Medicine. 2011, Bethesda, Maryland.

van der SLUIJS JP., PETERSEN AC., JANSSEN PHM., RISBEY JS, RAVETZ JR. Exploring the quality of evidence for complex and contested policy decisions. *Environ. Res.*

Lett. 2008 **3**(2).
WARDEKKER JA., van der SLUIJS JP., JANSSEN PHM., KLOPROGGE P., PETERSEN AC. Uncertainty communication in environmental assessments: views from the Dutch science-policy interface. *Environmental science & policy*. 2008 **11**: 627-641.
WITTERMAN C., RENOOIJ S. Evaluation of a verbal-numerical probability scale. *Intern Jourbal of Approximate Reasoning*. 2003 **33**: 117-131.