Deux approches pour catégoriser le risque

Natalia Grabar, Niña Kerry

STL UMR 8163 CNRS, Université Lille 3 et Lille 1
natalia.grabar@univ-lille3.fr, ninia@boujut.com
http://natalia.grabar.perso.sfr.fr/

Résumé. Le risque chimique ou alimentaire couvre les situations où les produits chimiques sont dangereux pour la santé et consommation humaine ou animale, et pour l'environnement. Les experts qui assurent le contrôle et la gestion de ces substances se retrouvent face à de gros volumes de littérature scientifique, qui doit être analysée pour appuyer la prise de décisions. Nous proposons une aide automatique pour l'analyse de cette littérature. Nous abordons la tâche comme une problématique de catégorisation: il s'agit de catégoriser les phrases des textes dans les classes du risque lié aux substances. Nous utilisons deux approches: par apprentissage supervisé et la recherche d'information. Les résultats obtenus avec l'apprentissage supervisé (toute classe confondue, F-mesure autour de 0,8 pour le risque alimentaire, entre 0,61 et 0,64 pour le risque chimique) sont meilleurs que ceux obtenus avec par recherche d'information (toute classe confondue, F-mesure entre 0,18 et 0,226 pour le risque alimentaire, entre 0,20 et 0,32 pour le risque chimique). Le rappel est compétitif avec les deux approches.

1 Introduction

Le risque chimique ou alimentaire se manifeste lorsque les produits chimiques sont dangereux pour la santé et consommation humaine ou animale, et pour l'environnement. Si certains produits et substances sont maintenant clairement identifiés comme dangereux (*e.g.* l'amiante, l'arsenic, le plomb), nos connaissances actuelles sur d'autres substances sont moins complètes. Nous nous intéressons en particulier au risque alimentaire (*e.g.* l'arsenic, les nitrates, la listeria, la dioxine) et au risque chimique (*e.g.* le bisphénol A, les phtalates). Ces substances entrent souvent dans la composition de produits courants et peuvent avoir l'effet nuisible sur l'organisme humain. Le contrôle sur la commercialisation de ces substances est effectué par des organismes sanitaires dédiés, comme EFSA (European Food Safety Authority) ou ANSES (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail). Les experts se retrouvent face à une littérature scientifique abondante et doivent l'étudier pour avoir une base solide pour la prise de décisions. L'objectif de notre travail consiste à proposer une aide automatique pour l'analyse de la littérature scientifique afin de détecter les phrases indicatives du risque induit par ces substances. Nous abordons cette tâche comme une problématique de catégorisation : les phrases des textes doivent être catégorisées dans les classes du risque. Nous présentons les données (section 2) et approches utilisées (sections 3 et 4). Nous discutons ensuite les résultats obtenus et concluons avec les pistes pour les travaux futurs (section 5).

2 Données utilisées dans les expériences

Notre objectif est de catégoriser les phrases des corpus dans les classes de risque. L'évaluation est effectuée par rapport aux données de référence. Une liste de mots vides et des ressources linguistiques sont aussi utilisées. Le travail est effectué avec le matériel en anglais.

Corpus. Les corpus proviennent de la littérature scientifique, qui est le matériel typique utilisé par les experts. Le corpus du risque chimique (80 000 occ.) contient le rapport sur le bisphénol A (EFSA Panel, 2010). Le corpus du risque alimentaire (>240 000 occ.) a été constitué à partir de 115 documents officiels publiés entre 2000 et 2010 sur une dizaine de substances, comme l'arsenic, la dioxine ou le nitrate (Blanchemanche et al., 2013). Trois sections (introduction, conclusion et résumé) sont traitées car elles comportent les résultats principaux.

Classifications du risque. Les classifications du risque (alimentaire (Blanchemanche et al., 2013) et chimique (Maxim et van der Sluijs, 2014)) sont structurées hiérarchiquement et décrivent différents aspects révélateurs de la nocivité des substances chimiques :

- significativité des résultats (*The Panel concluded that the current NOAEL for BPA (5 mg/kg b.w./day) would be sufficiently low to exclude any concern for this effect*) ;
- hypothèse scientifique (*Despite this lack of evidence, the possibility of poultry and egg consumption as an exposure route to HPAIV remains a concern to food safety experts*).

Le risque est présent lorsque la nocivité des substances est apparente dans la littérature scientifique, ou lorsque les expériences présentées montrent des imprécisions et incertitudes.

Ressources linguistiques. Des ressources linguistiques sont utilisées avec l'approche par apprentissage supervisé pour enrichir l'annotation. Nous supposons que ces différentes expressions, souvent liées à la notion d'incertitude, sont indicatrices de la notion du risque chimique :

- l'incertitude (*e.g. possible, should, may, usually*) indique des doutes existant au sujet des résultats obtenus expérimentalement, leur interprétation, etc. ;
- la négation (*e.g. no, neither, lack, absent, missing*) indique que de tels résultats n'ont pas été observés, que l'étude ne respecte pas les normes, etc. ;
- les limitations (*e.g. only, shortcoming, insufficient*) indiquent des limites, comme la taille insuffisante de l'échantillon traité, le faible nombre de tests ou de doses testées, etc. ;
- l'approximation (*e.g.*, *approximately, commonly, estimated*) indique d'autres insuffisances liées aux valeurs imprécises de substances, d'échantillons, de doses, etc.

Avec la recherche d'information, nous utilisons des ressources pour l'extension de requêtes :

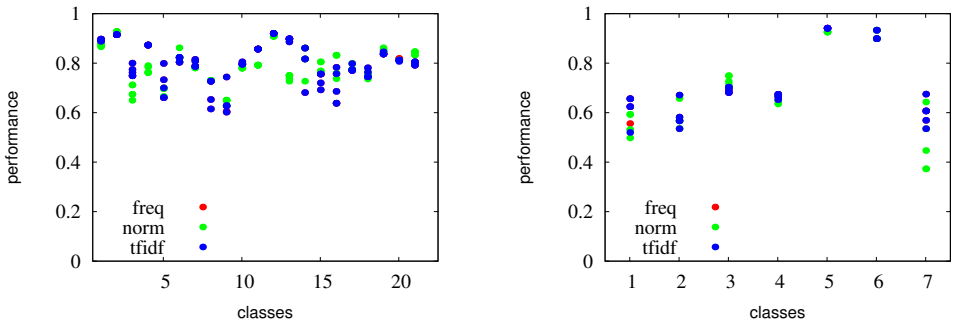
- 101 805 paires de synonymes provenant de la langue générale (Fellbaum, 1998) et spécialisée (Grabar et Hamon, 2010),
- des clusters de mots générés avec des méthodes distributionnelles à partir des corpus (Brown et al., 1992; Liang, 2005).

Données de référence. Les données de référence sont obtenues grâce à l'annotation par des spécialistes en évaluation du risque. Un expert a annoté 425 phrases couvrant 55 classes du risque chimique. Plusieurs experts ont participé dans l'annotation du corpus du risque alimentaire et fournissent des données de référence pour 657 phrases monoclasses couvrant 27 classes et 389 phrases multiclasses, pour un total de 1 046 phrases annotées. Plusieurs des classes contiennent très peu de phrases annotées et nous gardons celles qui fournissent un nombre suffisant d'exemples (minimum de 10 pour le risque alimentaire, 5 pour le risque chimique).

Liste de mots vides. La liste de mots vides contient 176 mots (*e.g.*, *& about again all almost and any by do to etc*). Cette liste contient essentiellement des mots grammaticaux.

3 Approche par apprentissage supervisé

Méthode. Nous utilisons différents algorithmes de la plateforme `Weka` (Witten et Frank, 2005) avec le paramétrage par défaut. Les phrases sont l'unité de travail. Nous visons la détection de phrases liées au risque : (1) de manière générale G pour détecter les phrases relatives au risque ; (2) de manière précise D pour associer ces phrases aux classes de risque. Les descripteurs sont fournis par l'annotation sémantique et linguistique : *forms* (les formes de mots comme elles apparaissent dans le corpus), *lemmas* (mots lemmatisés), *lf* (combinaison de formes et de lemmes), *tag* (les étiquettes morpho-syntaxiques des formes (*e.g.* noms, verbes, adjectifs)), *lft* (combinaison de formes, lemmes et étiquettes morpho-syntaxiques), *stag* (étiquettes sémantiques de mots (*e.g.* incertitude, négation, limitations)), *all* (combinaison de tous les descripteurs). Les descripteurs sont pondérés de trois manières : *freq* (fréquence brute des descripteurs), *norm* (fréquence normalisée par la taille du corpus), *tfidf* (pondération *tfidf* (Salton et Buckley, 1987)). Nous effectuons une validation croisée. Les mesures d'évaluation sont la précision, le rappel et la F-mesure (moyenne harmonique de la précision et du rappel).



(a) Risque alimentaire (b) Risque chimique

FIG. 1 – *Détection du risque par classe : F-mesure, lft, différentes pondérations.*

Résultats. Les résultats présentés sont obtenus avec J48 (Quinlan, 1993). Avec l'expérience G, les performances avec le risque alimentaire (autour de 0,8) sont meilleures que celles du risque chimique (0,61-0,64). Les performances sont assez stables avec les différents descripteurs et pondérations. L'exploitation de formes, d'étiquettes sémantiques et les différentes combinaisons de descripteurs donnent des résultats légèrement supérieurs. Bien que très simplistes, les étiquettes morpho-syntaxiques (*e.g.* noms, verbes, adjectifs) sont assez efficaces sur les deux corpus. Les étiquettes sémantiques seules (*stag*) sont parmi les plus efficaces pour détecter le risque chimique, mais montrent une F-mesure assez faible pour le risque alimentaire. À la figure 1, nous présentons l'expérience D avec les descripteurs *lft* (formes, lemmes et étiquettes morpho-syntaxiques). Les résultats sont élevés avec les classes du risque alimentaire et deux classes du risque chimique (*Facteur d'incertitude* et *Hypothèses scientifiques*). Le *tfidf* donne de meilleurs résultats dans la plupart des cas, mais la pondération *norm* est aussi compétitive. Les descripteurs *lft* fournissent de meilleurs résultats que les autres descripteurs. Les résultats sont meilleurs avec le risque alimentaire, où il existe plus de données d'apprentissage.

4 Approche de recherche d'information

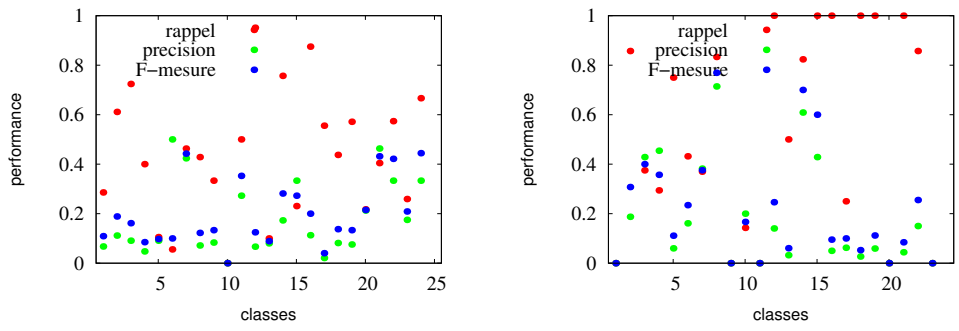
Méthode. Nous considérons les libellés des classes comme les requêtes et les phrases des corpus comme les réponses potentielles à ces requêtes. Nous exploitons le système de recherche d'information Indri (Strohman et al., 2005), qui utilise un modèle probabiliste basé sur le champ aléatoire de Markov et offre plusieurs fonctionnalités, comme par exemple :

- la racinisation (Porter (Porter, 1980) et de Krovetz (Krovetz, 1993)) réduit un mot à sa racine (*e.g.*, suppression de pluriels et de chaînes finales comme *-ment* et *-ique*) ;
- le *et* booléen (*band*) permet de combiner plusieurs mots clés ;
- les fenêtres ordonnées ou non ordonnées permettent de spécifier l'ordre des mots clés ;
- la pondération (*tfidf* (Salton et Buckley, 1987) et *okapi* (Robertson et al., 1998)) permet de relativiser le poids des mots-clés ;
- la pondération des synonymes (*wsyn*) permet d'indiquer l'importance des mots clés.

Pour l'expansion des requêtes, nous retenons les mots supplémentaires des ressources linguistiques (synonymes et clusters) si ces mots montrent au moins 0,3 % de précision. L'évaluation est effectuée avec plusieurs mesures : précision, rappel, F-mesure et MAP (Mean Average Precision), cette dernière prenant en compte l'ordre des réponses. Pour la *baseline*, les mots des libellés de classes sont utilisés, sans la racinisation ni l'expansion de requêtes.

	F-mesure		*MAP*	
	RA	*RC*	*RA*	*RC*
Baseline	0,18	0,20	0,13	0,21
Krovetz okapi	0,199	0,219	0,158	0,34
Krovetz tfidf	0,199	0,219	0,16	0,33
Porter okapi	0,191	0,20	0,156	0,289
Porter tfidf	0,191	0,20	0,157	0,277
Krovetz clusters sélection	0,226	0,32	0,142	0,26

TAB. 1 – *MAP : moyennes des performances avec les libellés des classes.*



(a) Risque alimentaire (b) Risque chimique

FIG. 2 – *Utilisation du raciniseur Krovetz et de la pondération* okapi.

Résultats. Le tableau 1 indique la MAP et la F-mesure de différentes expériences : *baseline*, utilisation de raciniseurs, pondération des mots clés et des clusters. Nous obtenons de meilleurs résultats avec les libellés du risque chimique, car ils sont plus explicites. La F-mesure est en général plus élevée que la MAP. Les résultats sont améliorés avec la racinisation, la pondération *tfidf* et *okapi*, et les clusters. Plusieurs autres expériences n'ont pas été concluantes (*e.g.* exploitation des définitions, pondération des synonymes, fenêtres ordonnées des mots clés des requêtes, *et* booléen). Krovetz, la pondération et les clusters fournissent les meilleurs résultats (figure 2). Les raciniseurs améliorent le rappel et donc les performances globales, tandis que l'utilisation de la pondération des mots clés (*okapi* ou *tfidf*) améliore surtout les valeurs de la MAP : les phrases retournées sont alors les mêmes, mais leur ordre devient plus correct. Les ressources linguistiques supplémentaires sont favorables pour certaines classes. Elles permettent surtout d'améliorer le rappel.

5 Discussion et Conclusion

L'apprentissage supervisé est plus performant que la recherche d'information, tandis que cette dernière, étant moins supervisée, permet de traiter un plus grand nombre de classes. La recherche d'information permet aussi de varier plus facilement les paramètres selon que l'on voudrait privilégier la précision ou le rappel. La pondération montre toujours un effet favorable. Dans une expérience similaire avec le risque alimentaire, des résultats comparables aux nôtres sont obtenus (Blanchemanche et al., 2013). Notons que nous avons aussi testé une approche non supervisée à base de règles, qui montre des résultats très faibles : rappel quasi-nul pour une précision entre 0,5 et 0,6. Il existe plusieurs possibilités pour combiner les deux approches testées : combinaison des sorties pour augmenter le rappel ; le vote des approches pour améliorer la précision ; l'utilisation des noeuds décisionnels des modèles d'apprentissage supervisé pour l'extension de requêtes ; l'exploitation des sorties de recherche d'information et du système à base de règles par l'apprentissage supervisé.

En conclusion, nous utilisons l'apprentissage supervisé et la recherche d'information pour détecter des phrases relatives au risque induit par les substances chimiques. Nous abordons la tâche comme une problématique de catégorisation : les phrases des textes doivent être catégorisées dans les classes de risque. Deux corpus et deux classifications du risque sont utilisés. Les résultats par apprentissage automatique sont les plus performants. Les résultats indiquent aussi que l'expression de l'incertitude linguistique (*e.g.*, *likely, should, assume*) est associée avec la notion du risque chimique. Dans les travaux futurs, nous allons tester d'autres paramètres pour améliorer les performances des approches testées et nous allons combiner les résultats de ces approches de différentes manières. Ces résultats peuvent être utilisés par les experts travaillant sur la gestion du risque pour la prise de décisions et évalués par eux.

Remerciements. Ce travail est soutenu par le projet PNRPE DICO-Risk.

Références

Blanchemanche, S., A. Rona-Tas, A. Duroy, et C. Martin (2013). Empirical ontology of scientific uncertainty : Expression of uncertainty in food risk analysis. In *Society for Social Studies of Science*, pp. 1–27.

Brown, P., P. deSouza, R. Mercer, V. Della Pietra, et J. Lai (1992). Class-based n-gram models of natural language. *Computational Linguistics 18*(4), 467–479.

EFSA Panel (2010). Scientific opinion on Bisphenol A : evaluation of a study investigating its neurodevelopmental toxicity, review of recent scientific literature on its toxicity and advice on the danish risk assessment of Bisphenol A. *EFSA journal 8*(9), 1–110.

Fellbaum, C. (1998). A semantic network of English : the mother of all WordNets. *Computers and Humanities. EuroWordNet : a multilingual database with lexical semantic network 32*(2-3), 209–220.

Grabar, N. et T. Hamon (2010). Exploitation of linguistic indicators for automatic weighting of synonyms induced within three biomedical terminologies. In *MEDINFO 2010*, pp. 1015–9.

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 191–202.

Liang, P. (2005). *Semi-Supervised Learning for Natural Language*. Master, Massachusetts Institute of Technology, Boston, USA.

Maxim, L. et J. P. van der Sluijs (2014). Qualichem in vivo : A tool for assessing the quality of in vivo studies and its application for Bisphenol A. *PLOS one*.

Porter, M. (1980). An algorithm for suffix stripping. *Program 14*(3), 130–137.

Quinlan, J. (1993). *C4.5 Programs for Machine Learning*. San Mateo, CA : Morgan Kaufmann.

Robertson, S., S. Walker, et M. Hancock-Beaulieu (1998). Okapi at TREC-7 : Automatic ad hoc, filtering, VLC and interactive. In *7th Text Retrieval Conference (TREC)*, pp. 199–210.

Salton, G. et C. Buckley (1987). Term weighting approaches in automatic text retrieval. Technical report, Department of computer science of Cornell university.

Strohman, T., D. Metzler, H. Turtle, et W. Croft (2005). Indri : a language-model based search engine for complex queries. In *International Conference on Intelligent Analysis*.

Witten, I. et E. Frank (2005). *Data mining : Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.

Summary

Chemical or food risk is relative to situations in which chemical products are dangerous for human or animal health and consumption, and for environment. The experts that guarantee the control and management of such substances face large amount of scientific literature, that have to be analyzed to support the decision making process. We propose an automatic assistance for the analysis of this literature. We tackle the task as the categorization problem: we want to categorize the sentences from corpora into classes of substance-related risk. We use two approaches: supervised machine learning and information retrieval. The results obtained with supervised machine learning (all classes together, F-measure around 0.8 for food risk, between 0.61 and 0.64 for chemical risk) are better than those obtained with information retrieval (all classes together, F-measure between 0.18 and 0.226 for food risk, between 0.20 and 0.32 for chemical risk). Recall is competitive with the two approaches.