Exploitation de différentes approches pour détecter et catégoriser le risque chimique et bactériologique

Natalia Grabar, Thierry Hamon

Exploitation de différentes approches pour détecter et catégoriser le risque chimique et bactériologique

Natalia Grabar[1] Thierry Hamon[2]
(1) CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France
(2) LIMSI-CNRS, BP133, Orsay; Université Paris 13, Sorbonne Paris Cité, France
natalia.grabar@univ-lille3.fr, hamon@limsi.fr

RÉSUMÉ

Le risque chimique couvre les situations où les produits chimiques sont ou peuvent être dangereux pour la santé humaine, animale et pour l'environnement. La détection des informations qui concernent le risque des substances chimiques occupe une place importante dans des agences environnementales et les chercheurs. 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. Notre objectif consiste 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. Nous utilisons trois approches : à base de règles, par apprentissage supervisé et la 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.

ABSTRACT

Exploitation of various approaches for automatic detection and categorization of chemical risk

Chemical risk corresponds to situations in which chemical products are or can be dangerous for human or animal health, or for environment. Detection of information on risk of chemicals occupies an important place in environmental agencies and researchers. Yet, large amounts of available data and controversies make it difficult to find the relevant information quickly and efficiently. Our objective is to propose an automatic help for the analysis of scientific literature in order to detect sentences indicative of chemical or bacteriological risk. We tackle the task as categorization problem: the sentences are are to be categorized in classes of risk. We use three approaches: rule-based, supervized categorization and information retrieval. The best results are obtained with supervized categorization and information retrieval. According to approaches, the results show up to 0.8 F-measure.

MOTS-CLÉS : Risque chimique, catégorisation supervisée, recherche d'information.

KEYWORDS: Chemical risk, supervised categorization, information retrieval.

1 Contexte

Le risque chimique couvre les situations où les produits chimiques sont ou peuvent être dangereux pour la santé humaine, animale et pour l'environnement. Le risque bactériologique concerne la contamination de l'alimentation par des bactéries ou produits chimiques. Comme les connaissances

actuelles sur certaines substances ne sont pas très complètes, cela laisse la possibilité d'avoir différents points de vue sur une substance donnée, ce qui peut mener à l'apparition de controverses. 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. De telles substances ont un impact grave pour la santé. La détection des informations qui indiquent leur risque occupe donc une place importante dans 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.

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 que présentent les substances chimiques. Nous abordons cette tâche comme une problématique de catégorisation : il s'agit de catégoriser les phrases des textes dans les classes du risque. Nous proposons d'utiliser trois approches : une approche à base de règles (section 3), une approche par apprentissage supervisé (section 4) et une approche exploitant un système de recherche d'information (section 5). Nous discutons ensuite les résultats obtenus et les comparons aux travaux existants (section 6). Avant de présenter nos expériences, nous introduisons les données communes à toutes ces expériences (section 2) : les classifications du risque, les corpus traités, les données de référence et une liste de mots vides.

2 Données communes aux expériences

Les expériences proposées ont pour objectif de traiter les phrases des corpus afin de les catégoriser dans les classes de risque. L'évaluation est effectuée par rapport aux données de référence. Une liste de mots vides est utilisée dans toutes les expériences proposées. Ce travail est mené sur l'anglais.

Corpus de travail. Les deux corpus traités proviennent de la littérature scientifique, qui est le matériel typique utilisé par les experts dans leur prise de décisions. Un premier corpus, traitant du risque chimique, contient le rapport sur le bisphénol A (EFSA Panel, 2010). Il comporte plus de 80 000 occurrences de mots. Un deuxième corpus, traitant du risque bactériologique, a été constitué à partir de 115 documents officiels produits entre 2000 et 2010 et portant sur une dizaine de substances ou de bactéries, comme l'arsenic, le dioxide, le nitrate ou la salmonelle (Blanchemanche *et al.*, 2013). Seuls trois sections sont considérées : introduction, conclusion et résumé, qui sont supposées contenir les informations nécessaires et suffisantes pour l'analyse du risque. Ce corpus comporte plus de 240 000 mots.

Classifications du risque. Les classifications hiérarchiques du risque décrivent différents aspects révélateurs de la nocivité des substances. Le risque est présent lorsque la nocivité des substances est apparente dans la littérature scientifique. De plus, le risque peut être présent lorsque les expériences présentées dans la littérature montrent des imprécisions et incertitudes relatives à différents aspects (*e.g.*, taille de l'échantillon, population, mesures, facteur de confusion, évaluation, extrapolation, présentation des résultats...). Les deux classifications exploitées (risque bactériologique (Blanchemanche *et al.*, 2009) et risque chimique (Maxim & van der Sluijs, 2014)) reflètent justement ces différents aspects du risque et des facteurs d'incertitude liés. La classification du risque bactéri-

ologique contient 28 classes, celle du risque chimique plus d'une centaine de classes. Chaque classe reçoit un libellé plus ou moins explicite de sa sémantique (*e.g.*, *missing factors/variables, inference from animal to human, limits of analytic methods, sample size, sampling method, natural variability*) et une définition.

Données de référence. Les données de référence sont obtenues grâce à l'annotation manuelle par des spécialistes en évaluation du risque. Un expert a annoté 425 phrases du risque chimique appartenant à 55 classes. Une phrase peut appartenir à plus d'une classe. Au moins deux experts ont annoté chaque phrase du risque bactériologique et fournis des données de référence consensuelles composées de 657 phrases monoclasses couvrant 27 classes et 389 phrases multiclasses, pour un total de 1 046 phrases annotées.

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 à base de règles

L'approche à base de règles repose sur une annotation sémantique des corpus et sur un ensemble de règles qui visent à mettre en relation les phrases des textes avec les classes.

3.1 Matériel supplémentaire

Des ressources linguistiques supplémentaires sont utilisées pour enrichir l'annotation :
- l'incertitude (Périnet *et al.*, 2011) (*e.g. possible, hypothetical, should, can, may, usually*), qui indique des doutes existant au sujet des résultats obtenus expérimentalement, leur interprétation, signification, etc. ;
- la négation (*e.g. no, not, neither, lack, absent, missing*), qui indique par exemple que de tels résultats n'ont pas été observés dans une étude, que l'étude ne respecte pas les normes, etc. Nous enrichissons une ressource existante (Chapman *et al.*, 2001) ;
- les limitations (*e.g. only, shortcoming, small, insufficient*), qui indiquent l'existence de limites, comme par exemple la taille insuffisante de l'échantillon traité, le faible nombre de tests effectués ou de doses testées, etc. ;
- l'approximation (*e.g.*, *approximately, commonly, considerably, estimated*), qui indique d'autres insuffisances liées à l'imprécision, comme par exemple des valeurs approximatives et imprécises de substances, d'échantillons, de doses, etc.

3.2 Méthode

Le traitement des corpus est fait en suivant plusieurs étapes :

1. Le pré-traitement est effectué avec la plateforme TAL Ogmios (Hamon & Nazarenko, 2008) et fournit les corpus et les libellés des classes normalisés linguistiquement grâce à leur segmentation, étiquetage morphosyntaxique et lemmatisation (Schmid, 1994) ;
2. L'annotation sémantique est obtenue suite à l'annotation des corpus avec les ressources linguistiques (section 3.1), afin de rendre les informations contenues dans les corpus plus explicites ;

3. Lors du post-traitement, nous nous basons sur l'intersection lexicale entre les libellés des classes et les phrases, ce qui garantit la pertinence du contenu des phrases, et le nombre de marqueurs (section 3.1), indicatif des incertitudes autour de ces classes de risque. Ainsi, dans la phrase *However, no specific measures were adopted to avoid sample contamination with free BPA during analytical procedures, which therefore cannot be excluded*, nous trouvons trois marqueurs (*however, no, cannot*), et tous les mots du libellé de la classe *Sample contamination*, ce qui permet de catégoriser la phrase dans cette classe. Nous testons plusieurs seuils pour ces deux valeurs (nombre de marqueurs et pourcentage de mots d'une classe donnée) ;
4. L'évaluation est effectuée par rapport aux données de référence avec deux mesures : précision (pourcentage de phrases correctes parmi toutes les phrases trouvées) et rappel (pourcentage de phrases correctes parmi toutes les phrases attendues).

3.3 Résultats de l'approche à base de règles

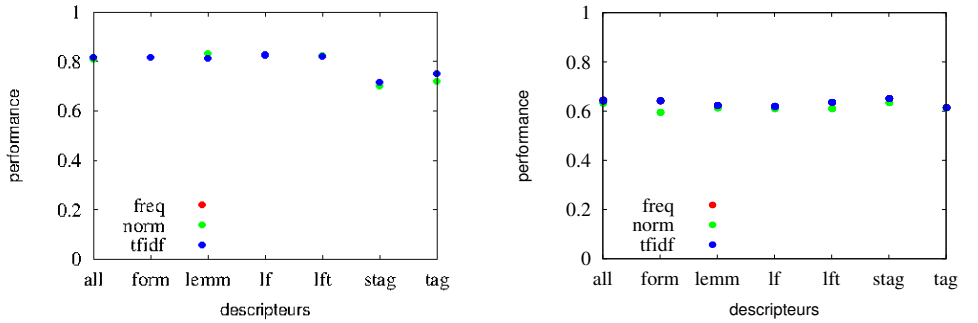
Les résultats montrent qu'avec l'augmentation de contraintes (nombre de marqueurs et pourcentage de mots communs dans les libellés et le texte) le nombre de phrases extraites diminue tandis que la précision augmente. Sur le corpus du risque chimique, les meilleurs seuils observés sont entre 55 % et 60 % de mots communs : le nombre de phrases est alors le plus important, tandis que la précision devient acceptable (0,5-0,65). Les marqueurs de l'incertitude sont toujours nécessaires pour la détection du risque. Toutefois, l'appariement direct entre les libellés des classes et le texte ne suffit pas : le rappel reste extrêmement bas (moins de 0,1). Parmi les 34 classes testées, la méthode permet d'extraire les phrases pour 18 classes. Une analyse manuelle a montré que la méthode détecte aussi des phrases correctes qui ne font pas partie des données de référence. Si ces phrases sont prises en compte, la précision augmente. Parmi les limites, dans plusieurs phrases extraites, il n'existe pas de lien syntaxique ni sémantique entre les marqueurs et les mots des libellés. Pour corriger cet aspect, une analyse syntaxique ou bien la fenêtre graphique délimitée peuvent être utilisées. Comme souligné, la méthode souffre surtout du manque de couverture. Cela arrive lorsqu'il n'existe pas de correspondance directe entre les libellés et les mots utilisés dans les corpus, comme *GLP compliance* et *GLP compliant*, ou *dose* et *dosage* : un lexique spécifique de synonymes doit être utilisé. Par ailleurs, certains libellés ne sont pas évocateurs de leur contenu. Étant donné les faibles performances de cette approche, nous nous orientons vers d'autres approches, mais nous retenons que les marqueurs de l'incertitude s'avèrent utiles pour la tâche.

4 Approche par apprentissage supervisé

L'approche par apprentissage supervisé exploite les corpus avec les mêmes annotations que celles de la section 3 : les corpus sont annotés linguistiquement (étiquetage morpho-syntaxique, lemmatisation) et sémantiquement (incertitude, négation, limitations et approximation).

4.1 Méthode

L'approche par apprentissage supervisé nécessite un ensemble d'entraînement pour générer les modèles qui peuvent être appliqués aux nouvelles données. Comme auparavant, les phrases sont



(a) Risque bactériologique (b) Risque chimique

Fig. 1 – F-mesure obtenue avec la tâche G : détection de phrases liées au risque.

notre unité de travail. Nous utilisons différents algorithmes de la plateforme `Weka` (Witten & Frank, 2005) avec le paramétrage par défaut. Nous visons la détection de phrases liées au risque à deux niveaux : (1) de manière générale (expérience G), où il s'agit de détecter les phrases pertinentes sans les associer avec les classes de risque ; (2) de manière plus précise (expérience D), où il s'agit de détecter les phrases pertinentes et de les associer aux classes de risque. Pour chaque expérience et classe, les données à traiter sont sélectionnées de manière équilibrée, avec un nombre égal de phrases pertinentes et non pertinentes. Par exemple, lorsque les phrases sont à catégoriser au niveau général, nous utilisons les 425 phrases annotées avec le risque chimique et 425 de phrases non annotées du même corpus. Pour les tests avec les classes individuelles, nous suivons le même principe. Lors du recrutement des phrases non pertinentes pour les classes individuelles, nous combinons les phrases non annotées et les phrases annotées par d'autres classes. Les descripteurs utilisés 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.

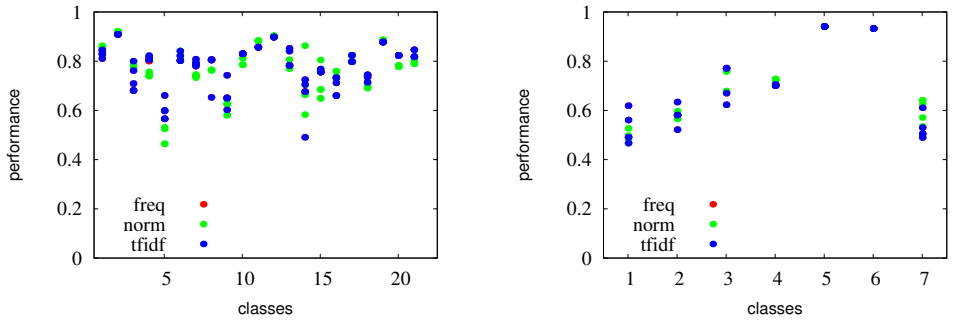
Les descripteurs sont pondérés de trois manières :

– *freq* correspond à la fréquence brute des descripteurs ;
– *norm* correspond à la fréquence normalisée par la taille du corpus (en nombre de mots) ;
– *tfidf* correspond à la pondération *tfidf* des fréquences brutes (Salton & Buckley, 1987).

Nous effectuons une validation croisée (Sebastiani, 2002) avec trois mesures d'évaluation (précision, rappel, F-mesure). La *baseline* correspond à l'assignation des phrases dans la catégorie par défaut.

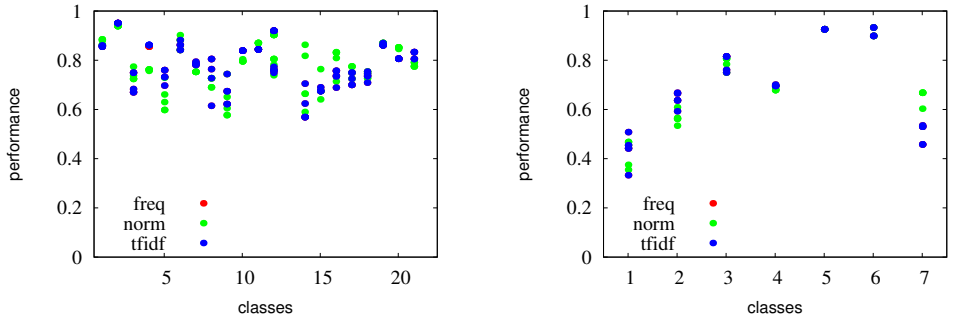
4.2 Résultats de l'approche par apprentissage supervisé

Les résultats sont présentés et discutés en fonction de deux tâches : G détection de phrases liées au risque et D détection de phrases pertinentes et leur assignation dans les catégories du risque. Nous présentons ici les résultats obtenus avec J48 (Quinlan, 1993).



(a) Risque bactériologique (b) Risque chimique

Fig. 2 – Détection du risque par classe : F-mesure, *formes*, différentes pondérations.

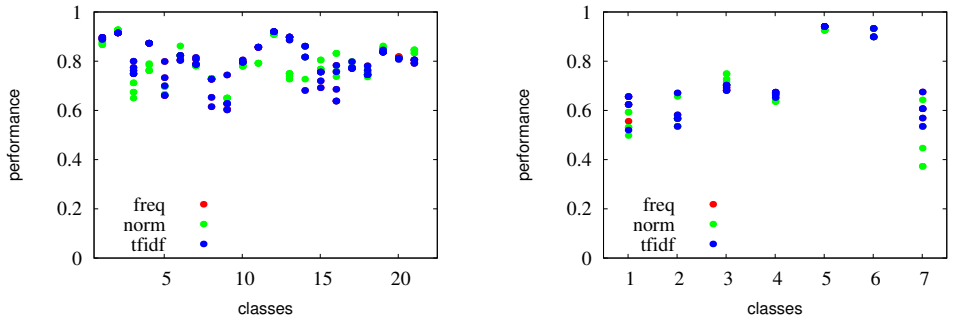


(a) Risque bactériologique (b) Risque chimique

Fig. 3 – Détection du risque par classe : F-mesure, *lemmes*, différentes pondérations.

La figure 1 indique les résultats pour la tâche G : risque bactériologique (en 1(a)) et risque chimique (en 1(b)). L'axe x indique les descripteurs utilisés (formes, lemmes, étiquettes sémantiques et morpho-syntaxiques et différentes combinaisons). L'axe y indique la F-mesure. Les couleurs des points correspondent à la pondération des descripteurs (fréquence brute, normalisée ou pondérée par $tfidf$). Nous observons que, globalement, les performances avec le risque bactériologique (autour de 0,8) sont meilleures que celles avec le risque chimique (entre 0,61 et 0,64). Elles sont stables avec les différents descripteurs et pondérations. L'exploitation de formes, d'étiquettes sémantiques et les différentes combinaisons donnent des résultats légèrement supérieurs. Bien que très simplistes, les étiquettes morpho-syntaxiques sont efficaces. Les étiquettes sémantiques seules sont parmi les plus efficaces avec le risque chimique, mais montrent une F-mesure faible avec le risque bactériologique.

Les figures 2 à 4 présentent les résultats pour les classes individuelles, en utilisant comme descripteurs les formes, les lemmes et la combinaison *lft*. 22 classes du risque bactériologique et 7 classes du risque chimique peuvent être traitées car elles reçoivent un nombre suffisant de phrases. Nous retenons les classes ayant au moins 10 phrases pour le risque bactériologique et au moins 5 phrases pour le risque chimique, ce deuxième corpus proposant en effet moins de phrases annotées.



(a) Risque bactériologique (b) Risque chimique

Fig. 4 – Détection du risque par classe : F-mesure, *lft* (lemmes/formes/étiquettes morpho-syntaxiques), différentes pondérations.

Les figures indiquent que les résultats sont meilleurs avec le risque bactériologique et deux classes du risque chimique (*Choix du facteur d'incertitude*, *Hypothèses scientifiques*). La pondération influence les résultats : *tfidf* est le meilleur, *norm* reste compétitif. Concernant les trois types de descripteurs, ils sont équivalents entre eux, tandis que la variation de la F-mesure est moins importante avec les descripteurs *lft*. C'est aussi avec *lft* que nous obtenons de meilleurs résultats. Nous voyons que les résultats sont meilleurs avec les classes du risque bactériologique, pour lesquelles il existe plus de données d'entraînement. Cependant, pour certaines classes (*e.g. Choix du facteur d'incertitude*, *Variabilité naturelle et non expliquée*), il n'existe pas de relations entre la taille et les performances.

5 Approche de recherche d'information

Avec l'approche de recherche d'information, nous considérons les libellés des classes comme les requêtes et les phrases des corpus comme les réponses. Les corpus bruts, les libellés des classes et les définitions des classes sont exploités avec un système de recherche d'information.

5.1 Matériel supplémentaire

Nous utilisons des ressources supplémentaires : 101 805 paires de synonymes provenant de la langue générale (Fellbaum, 1998) et spécialisée (Grabar & Hamon, 2010) et des clusters de mots générés de manière non supervisée à partir des corpus avec des méthodes distributionnelles (Brown *et al.*, 1992; Liang, 2005). Lors de la génération de clusters, nous générons entre 200 et 600 clusters.

5.2 Méthode

Nous exploitons le système de recherche d'information Indri (Strohman *et al.*, 2005), qui montre de bonnes performances dans la littérature. Ce système utilise un modèle probabiliste basé sur le champ aléatoire de Markov. Nous utilisons plusieurs fonctionnalités proposées par Indri :

– Racinisation, qui permet de ramener un mot à sa racine grâce à la suppression de finales de mots (*e.g.*, pluriels, *-ment*, *-ique*) et d'augmenter ainsi le rappel. Nous effectuons des expériences sans le raciniseur, et avec les raciniseurs de Porter (Porter, 1980) et de Krovetz (Krovetz, 1993) ;
– Utilisation de *et* booléen (*band*), qui permet de combiner plusieurs mots clés ;
– Fenêtres ordonnées ou non ordonnées, qui permettent de spécifier l'ordre des mots clés ;
– Pondération des mots clés, qui permet de relativiser leur poids par rapport à la phrase ou le corpus. Nous utilisons deux mesures : *tfidf* (Salton & Buckley, 1987) et *okapi* (Robertson *et al.*, 1998; Claveau, 2012), cette dernière étant conseillée lors du traitement des documents courts ;
– Pondération des synonymes (*wsyn*), qui permet d'indiquer l'importance des mots clés.

Pour l'expansion des requêtes, nous retenons les mots provenant de ressources supplémentaires (section 5.1) si ces mots clés montrent au moins 0,3 % de précision. Chaque mot clé peut ainsi être étendu avec ses synonymes ou les mots de ses clusters. 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*, seuls les mots des libellés de classes (ou de leurs définitions) sont utilisés, sans racinisation ni expansion de requêtes.

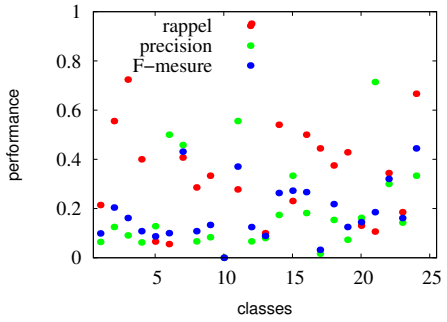
5.3 Résultats de l'approche par recherche d'information

Nous pouvons traiter 24 classes du risque bactériologique et 23 classes du risque chimique. Le tableau 1 indique les moyennes de MAP et de F-mesure : *baseline*, requêtes avec les raciniseurs, pondération des mots clés et des clusters générés de manière non supervisée (section 5.1). De manière générale, nous obtenons de meilleurs résultats avec le risque chimique. Nous pensons que leurs libellés sont plus explicites quant à leur contenu. La F-mesure est plus élevée que la MAP. Ces expériences montrent une amélioration systématique par rapport à la *baseline*. Il s'agit de l'utilisation des algorithmes de racinisation Porter et Krovetz, ce dernier étant souvent meilleur ; de la pondération des mots clés avec *tfidf* et *okapi* ; et de l'utilisation des clusters avec la sélection de mots clés. Plusieurs autres expériences testées n'ont pas été concluantes (*e.g.* exploitation des définitions, pondération des synonymes, fenêtres ordonnées des mots clés des requêtes, utilisation de *et* booléen). Le raciniseur de Krovetz avec la pondération et les clusters fournissent les meilleurs résultats.

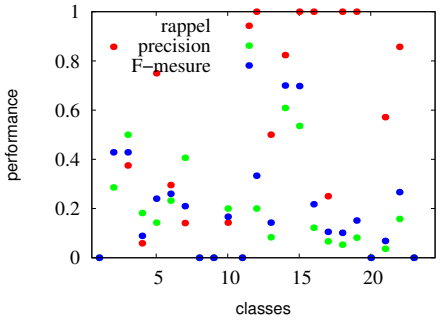
	F-mesure		*MAP*	
	RA	*RC*	*RA*	*RC*
Baseline	0,18	0,20	0,13	0,21
Krovetz okapi	0,199	0,219	0,158	0,34
Krovetz tfidf	0,199	0,219	0,16	0,33
Porter okapi	0,191	0,20	0,156	0,289
Porter tfidf	0,191	0,20	0,157	0,277
Krovetz clusters sélection	0,226	0,32	0,142	0,26

TABLE 1 – Différentes expériences : moyennes avec les libellés des classes (MAP et F-mesure).

Concernant les résultats par classe de risque, à la figure 5, nous présentons les résultats obtenus avec la baseline. L'axe x indique les classes, l'axe y les performances obtenues en termes de précision (vert), rappel (rouge) et F-mesure (bleu). Nous pouvons observer que les résultats sont assez variables selon les classes et que certaines classes reçoivent pas ou peu de réponses. Comme indiqué, les résultats sont meilleurs avec le corpus du risque chimique. À la figure 6, nous présentons les résultats lorsque le raciniseur Krovetz et la pondération *okapi* sont utilisés. Comme nous l'avons indiqué, l'application

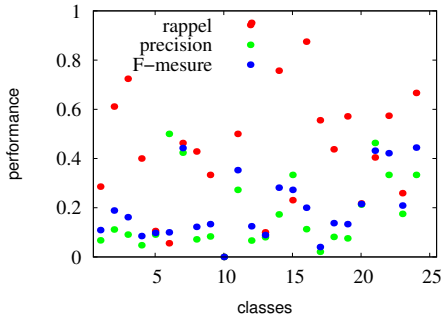


(a) Risque bactériologique

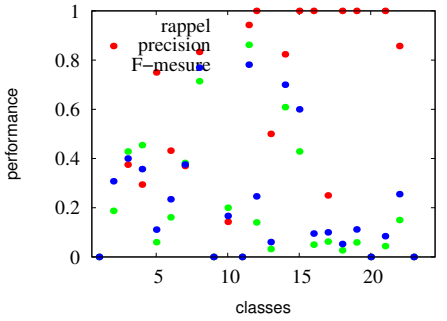


(b) Risque chimique

Fig. 5 – Baseline avec la recherche d'information : performances par classe.



(a) Risque bactériologique



(b) Risque chimique

Fig. 6 – Utilisation du raciniseur Krovetz et de la pondération *okapi*.

des raciniseurs améliore le rappel et donc les performances globales, tandis que l'utilisation de la pondération des mots clés (*okapi* ou *tfidf*) améliore surtout la MAP. Dans ce dernier cas, les phrases retournées sont les mêmes, mais leur ordre devient plus correct. Les valeurs de la précision peuvent diminuer pour certaines classes. En revanche, les performances globales sont améliorées de plusieurs points. Par ailleurs, avec l'utilisation de ressources linguistiques supplémentaires, l'impact varie en fonction des classes : il est positif pour certaines classes 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.

6 Comparaison des approches

Une comparaison entre les trois approches testées (à base de règles, par apprentissage supervisé et par recherche d'information) montre que l'apprentissage supervisé et la recherche d'information sont

plus performants que l'approche à base de règles. Cette dernière montre effectivement un rappel extrêmement faible pour une précision autour de 0,5 à 0,6. L'approche par la recherche d'information permet de traiter un plus grand nombre de classes, tandis que l'apprentissage supervisé permet d'obtenir de meilleurs résultats pour moins de classes. Cependant, avec l'apprentissage supervisé, la taille des données de référence doit être plus importante, même si nous avons observé que pour certaines classes, qui montrent des spécificités de contenu, une petite taille de données peut être suffisante. L'approche par la recherche d'information permet de varier plus facilement les paramètres selon que l'on voudrait privilégier la précision ou le rappel. Pour ces deux approches, la pondération des descripteurs et des mots clés montre toujours un effet favorable. Dans une expérience similaire sur les données du risque bactériologique, et utilisant une approche par apprentissage automatique, des résultats comparables aux nôtres (section 4.2) ont été obtenus (Blanchemanche *et al.*, 2013).

Il existe plusieurs possibilités pour combiner les trois approches. Cette combinaison peut permettre de détecter plus de phrases pertinentes (amélioration du rappel) et d'améliorer ainsi les performances globales. Le vote de ces approches et la sélection des décisions majoritaires peut aussi être effectué si c'est l'amélioration de la précision qui est visée. Nous pouvons aussi utiliser les noeuds décisionnels des modèles fournis par l'apprentissage supervisé pour fournir d'autres alternatives à l'extension de requêtes. Finalement, le chaînage des approches est aussi possible : l'approche par l'apprentissage supervisé peut exploiter les descripteurs proposés avec les sorties des deux autres approches.

7 Conclusion et Perspectives

Nous avons présenté les expériences effectuées avec trois approches (à base de règles, par apprentissage supervisé et par recherche d'information) appliquées à la tâche de détection de phrases relatives au risque induit par les substances chimiques ou par les bactéries, en relation avec le risque chimique ou bactériologique, respectivement. 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 : ceux dédiés au risque bactériologique et ceux dédiés au 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. Les résultats par apprentissage automatique sont les plus performants : cette approche permet de traiter le plus grand nombre de classes et de phrases, et d'obtenir les résultats plus performants. Dans les travaux futurs, nous allons tester d'autres paramètres pour améliorer les performances des approches testées. Par ailleurs, nous pouvons aussi combiner les résultats de ces approches, en effectuant la fusion des phrases catégorisées ou leur vote. Nous pouvons aussi utiliser les noeuds décisionnels des modèles fournis par l'apprentissage supervisé pour obtenir des ressources alternatives ou supplémentaires à l'extension de requêtes. Finalement, les résultats générés peuvent être utilisés par les experts travaillant sur la gestion du risque pour la prise de décisions. De la même manière, cela pourra fournir une évaluation supplémentaire des résultats.

Références

BLANCHEMANCHE S., BUCHE P., DIBIE-BARTHÉLÉMY J., MÉLÈZE E. F., IBANESCU L. & RONA-TAS A. (2009). Ontology building: an application in food risk analysis. In *TIA*.

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.

BROWN P., DESOUZA P., MERCER R., DELLA PIETRA V. & LAI J. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, **18**(4), 467–479.

CHAPMAN W., BRIDEWELL W., HANBURY P., COOPER G. & BUCHANAN B. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform. 2001 Oct ;34(5):*, **34**(5), 301–10.

CLAVEAU V. (2012). Vectorisation, okapi et calcul de similarité pour le tal : pour oublier enfin le tf-idf. In *Traitement Automatique des Langues Naturelles (TALN)*, p. 85–98.

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. & HAMON T. (2010). Exploitation of linguistic indicators for automatic weighting of synonyms induced within three biomedical terminologies. In *MEDINFO 2010*, p. 1015–9.

HAMON T. & NAZARENKO A. (2008). Le développement d'une plate-forme pour l'annotation spécialisée de documents web: retour d'expérience. *TAL*, **49**(2), 127–154.

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*, p. 191–202.

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

MAXIM L. & VAN DER SLUIJS J. P. (2014). Qualichem in vivo: A tool for assessing the quality of in vivo studies and its application for Bisphenol A. *PLOS one*.

PÉRINET A., GRABAR N. & HAMON T. (2011). Identification des assertions dans les textes médicaux: application à la relation {patient, problème médical}. *TAL*, **52**(1), 97–132.

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., WALKER S. & HANCOCK-BEAULIEU M. (1998). Okapi at TREC-7 : Automatic ad hoc, filtering, VLC and interactive. In *7th Text Retrieval Conference (TREC)*, p. 199–210.

SALTON G. & BUCKLEY C. (1987). *Term weighting approaches in automatic text retrieval*. Rapport interne, Department of computer science of Cornell university.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *ICNMLP*, p. 44–49, Manchester, UK.

SEBASTIANI F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1), 1–47.

STROHMAN T., METZLER D., TURTLE H. & CROFT W. (2005). Indri: a language-model based search engine for complex queries. In *International Conference on Intelligent Analysis*.

WITTEN I. & FRANK E. (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.