Machine learning-based detection of chemical risk

Natalia GRABAR[a,1] and Ornella WANDJI TCHAMI[a] and Laura MAXIM[b]
[a]*CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France*
[b]*Institut des Sciences de la Communication, CNRS UPS 3088, France*

Abstract. Chemical risk appears with chemical substances that are dangerous for human or animal health, or for environment, such as with Bisphenol A and phtalates. Chemical risk causes several severe health disorders and is particularly dangerous for human health. Specific agencies are involved in the verification of the suitability of products and goods to be marketed. For this, a large amount of scientific and institutional literature is manually analyzed to study the current knowledge on the associated chemical risk. We propose to use machine learning and dedicated classification for the automatic detection of chemical risk statements. We test several algorithms and features and obtain between 0.60 and 0.95 F-measure.

Keywords. Natural Language Processing, Public Health, Machine Learning, Chemical Risk, Uncertainty, Evaluation

Introduction

Chemical risk appears when chemical substances, dangerous for human or animal health or for environment, are used. For instance, substances like Bisphenol A and phtalates are endocrine disrupters that, at certain doses, can interfere with the endocrine (or hormone system) in mammals [1,2]. They cause a great number of severe disorders such as sexual development problems (feminizing of males or masculine effects on females), breast cancer, prostate cancer, thyroid and other cancers, brain development problems and deformations of the body [3]. Although controversial, the effect of endocrine disrupters occupies an important place in medical and economical spheres. For these reasons, authorization for marketing of some products and goods may depend on their chemical composition. Such authorizations are delivered by specific agencies like European Food Safety Authority (EFSA) [4]. People working in these agencies have to analyze a great amount of literature to provide scientifically-based arguments for decision-makers on possibility and appropriateness of marketing of these products and goods. We propose an automatic approach for the analysis of literature and for detection of sentences related to chemical risk. In the following of this paper, we present first the material used, we then introduce the methods, discuss the results obtained and conclude with some future work.

1 Corresponding Author.

1. Material

Processed corpus. The corpus processed [5] proposes review of scientific literature on Bisphenol A-related experiments and known experimental results. It contains over 80,000 word occurrences. The corpus contains typical statements that may be found in scientific and institutional literature. The arguments provided are used for supporting the decisions for managing the chemical risk uncertainties.

Linguistic resources. Linguistic resources contain linguistic *markers* for describing several aspects of the content of scientific writings: *negation* [6] (*eg, no, not, neither, lack, absent, missing*), which indicate that experimental results have not been observed in a study, that a study did not respect the norms, etc.; *uncertainty* [7] (*eg, possible, hypothetical, should, can, may, usually*), which indicates doubts about the results, their interpretation, significance, etc.; *limitations* (*eg, only, shortcoming, small, insufficient*), which indicate limits, such as small size of a sample, small number of tests or doses, etc.; *approximation* (*eg, approximately, commonly, considerably, estimated*), which indicates limits due to an imprecise or approximate values of substances, samples, doses, etc. Such markers are frequent in the scientific writings, especially negation and uncertainty that have been quite well studied in previous work [8,9]. The limitation and approximation markers do not seem to have been used in existing work.

Classification. The classification describes types of chemical risk factors and uncertainties like causal relationship between the chemicals and the induced risk, such as laboratory procedures, human factors, animals tested, significance of results, form of reporting, natural variability, control of confounders, exposure, dosage, assumptions, performance of the measurement and analytical method, etc. [10].

Reference data. The reference data are obtained thanks to a manual annotation by a specialist of chemical risk: 425 segments are assigned to 55 classes of risk factors. The classes to not overlap. The reference data are a subset of the whole corpus.

2. Methods

Machine learning, or supervised categorization, is exploited for performing the automatic detection of sentences that mention chemical risk uncertainties and different classes of the chemical risk uncertainties.

Pre-processing and Annotation. Corpus is pre-processed with the Ogmios plateform [11] through tokenization into words, POS-tagging and lemmatization by Genia tagger [12]. The corpus is also annotated with the linguistic resources.

Supervised categorization. For the automatic detection of risk factors, we use the reference data and supervised categorization [13]. The categories to be recognized are positioned at two levels: first, we want to detect sentences concerned with the chemical risk uncertainties; second, we want to detect to which classes of chemical risk these sentences belong. In the first case, we address a more general problem, while in the second case, we try to perform a fine-grained classification of sentences. For each experiment, we build datasets with equal numbers of positive and negative examples. For instance, when sentences are to be categorized at the level of chemical risk uncertainties, we recruit the 425 positive sentences related to chemical risk and 425 negative sentences non-related to chemical risk. We proceed similarly when recruiting

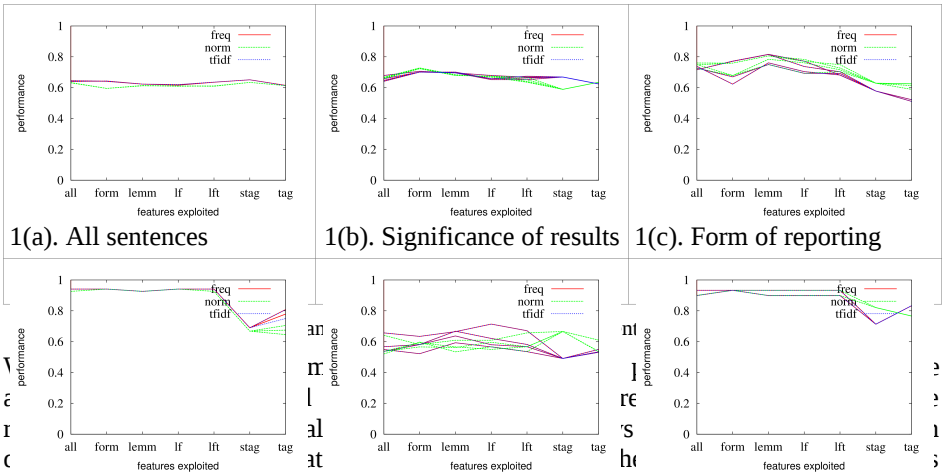
data for different classes, but in this case, the negative set combines both sentences that are non-related to chemical risk and sentences that are related to other classes of risk.

Features used. The features used for the categorization are issued from the semantic annotation: *forms* (words as they occur in the texts: *uncertain*, *risks*); *lemmas* (lemmatized words: *uncertain*, *risk*); *lf* (combination of forms and lemmas: *uncertain/uncertain*, *risks/risk*); *tag* (morpho-syntactic or POS tags of the words: *adj*, *noun*); *lft* (combination of forms, lemmas and tags: *uncertain/uncertain/adj*, *risks/risk/noun*); *stag* (semantic tags of the words, such as *uncertainty*, *negation*, *limitations*, *approximation*); *all* (combination of all the features). The features are selected according to the number of documents they occur within. We test several values, going from 1 to 4 documents. The objective is to reduce the decision space and to select the most relevant features.

Feature weighting. These features are weighted in three ways: *freq* (row frequency of features); *norm* (normalization of the frequency by the document length); *tfidf* (weighting of the frequency by term frequency*inverse document frequency [14]).

Evaluation and Baseline. We apply the cross-validation [15], designed to use two distinct data sets during training and test steps. We perform a three-fold cross-validation. We compute three evaluation measures: precision (percentage of sentences that are correctly categorized among all the sentences categorized in a given category); recall (percentage of sentences correctly categorized comparing to those that should be categorized in a given category); f-measure (their harmonic mean). We also exploit a *baseline*, that corresponds to the assignment of sentences in the default category. For instance, with a two-category test (*eg* sentences related to the chemical risk uncertainties and sentences non-related to the chemical risk uncertainties) and the equal number of sentences in each category, this kind of *baseline* would generated 50% performance, as all sentences would be assigned to the same category. Comparing to this *baseline*, we also compute the obtained gain, that corresponds to the real improvement of the performance *P* by comparison with the *baseline BL* [16]: *(P-BL)/ (1-BL).*

3. Results and Discussion



1(a). All sentences 1(b). Significance of results 1(c). Form of reporting

indicates the features used (forms, lemmas, semantic tags, POS tags or their different combinations). The y-axis indicates the performance values (F-measure). Different colors of the lines correspond to the weighting of the features (row frequency, normalization by the document length, normalization with the tfidf). Moreover, for each normalization, different thresholds for the feature selection are applied. Figure 1(a) indicates that the performance is stable (between 0.61 and 0.64), with different features used. Exploitation of forms, semantic tags and combination of all the features give slightly better results. Although very simplistic, the POS tags *(eg* nouns, verbs, adjectives) appear to be quite efficient to the task. Concerning other settings of the method: we obtain similar results with different normalization methods (row frequency, normalization by the document length, normalization with the tfidf); the feature selection has no impact of the results. Compared to the baseline (0.50 F-measure), we obtain 0.24 points of improvement.

Classes that contain at least ten sentences are treated. We present here the results obtained on five classes: *Significance of results* (Figure 1(b)), *Form of reporting* (Figure 1(c)), *Natural variability* (Figure 1(d)), *Control of confounders* (Figure 1(e)) and *Assumptions* (Figure 1(f)). We can observe that there is no direct and observable dependance between the size of the datasets and the performance obtained: the results are similar for several classes whichever the feature selection and weighting, while with other classes the variability of the results is important. Similarly, the results may depend on the feature selection, especially with the class *Confounder control*. In general, we can observe poorer performance with semantic tags (stag) and POS tags (tag). Usually, forms, lemmas and their combinations with other features appear to be efficient. Compared to the baseline, the highest gain is obtained with the class *Natural variability*, which means that this class shows specific linguistic features (lexicon, markers) by comparison with the rest of the sentences. The automatic detection is then facilitated by this fact. Here is an example of sentence related to the class *Natural variability*: "*Based on the re-analysis the Panel considered that no conclusion can be drawn from this study on the effect of BPA on learning and memory behaviour due to large variability in the data.*" Among the salient features, we can find words such as *variability*, *differences* for the class *Natural variability*; *factor, uncertainty* for the class *Choice of uncertainty factors*; *contamination*, *not*, *confounders*, *confounding* for the class *Control of confounders*. They are often evocative of the semantics of the corresponding class.

4. Conclusion and Perspectives

The presented work provide experimental insights into the automatic detection of sentences related to the chemical risk uncertainties. We proposed to perform the extraction of sentences concerned by chemical risk at two levels: detection of sentences potentially concerned by the chemical risk uncertainties; detection of sentences potentially concerned by specific classes of chemical risk uncertainties. Several classes from the reference data contain but a small number of sentences: their automatic categorization is not performed currently. One possible solution is to perform the over-sampling [18]. Still, with the current experiments, the limitation due to the size of classes is partially overcome by the fact that sentences under these classes can be detected without specifying their category *(ie*, general relation to the chemical risk

uncertainties and not to a given class of chemical risk uncertainties). The performance obtained is respectable: it varies between 0.60-0.70 for classes that are difficult to detect, and up to 0.82-0.95 for classes that show lexical and semantic specificities.

Our future work will address the current limitations of the presented experiments: building the dedicated lexicon, application of over-sampling algorithms, use of other methods (topic modeling, information retrieval). Besides, the proposed methods will be applied to a larger corpus of documents from the chemical risk area, as well as corpora concerned with other chemical substances. The extracted results will be presented for evaluation and assessment to experts working in environmental agencies.

5. References

[1] Testai E, Galli CL, Dekant W, Marinovich M, Piersma AH, Sharpe RM. A plea for risk assessment of endocrine disrupting chemicals. *Toxicology*. 2013;**314**(1), 51-9.

[2] Andersson AM, Bay K, Grigor KM, Rajpert-De Meyts E, Skakkebaek NE. Special issue on the impact of endocrine disrupters on reproductive health. *Int J Androl*. 2012;**35**(3), 215.

[3] Crisp TM, Clegg ED, Cooper RL, Wood WP, Anderson DG, Baetcke KP, Hoffmann JL, Morrow MS, Rodier DJ, Schaeffer JE, Touart LW, Zeeman MG, Patel YM (1998). Environmental endocrine disruption: An effects assessment and analysis. *Environ. Health Perspect*. **106.** (Suppl 1), 11–56.

[4] www.efsa.europa.eu/

[5] 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. *European Food Safety Authority (EFSA) journal*, **8**(9), 1-110.

[6] Chapman, W., Bridewell, W., Hanbury, P., Cooper, G., Buchanan, B. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. 2001, **34**(5), 301-10.

[7] Périnet, A., Grabar, N., Hamon, T. Identification des assertions dans les textes médicaux: application à la relation {patient, problème médical}. *TAL* 2011 **52**(1), 97-132.

[8] Mutalik, P., Deshpande, A., Nadkarni, P. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *J Am Med Inform Assoc 2001*, **8**(6), 598-609.

[9] Light, M., Qiu, XY., Srinivasan, P. The language of bioscience: facts, speculations and statements in between. In *ACL WS on Linking biological literature, ontologies and databases 2004:* 17-24.

[10] Maxim L and 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. In press

[11] Hamon, T., Nazarenko. Le développement d'une plate-forme pour l'annotation spécialisée de documents web: retour d'expérience. *TAL 2008*, **49(2),** 127-154.

[12] Tsuruoka, Y., Tateishi, Y., Kim, J., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J. Developing a robust part-of-speech tagger for biomedical text. *LNCS 2005*, **3746,** 382-392.

[13] Witten, I., Frank, E. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.

[14] Salton, G. Developments in automatic text retrieval. *Science 1991*, **253,** 974-979.

[15] Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Surveys 2002*, **34**(1)**,** 1-47.

[16] Rittman, R. *Automatic discrimination of genres.* Saarbrucken, Germany: VDM, 2008.

[17] Quinlan, J. *C4.5 Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[18] Chawla, N.V., Bowyer, K.W., Hall, L. O., Kegelmeyer, W.P. Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 2002*, **16,** 321-357.