Descriptors for the detection of the chemical risk

Natalia Grabar
UMR8163 STL
CNRS, Université Lille 3
Villeneuve d'Ascq, France
natalia.grabar@univ-lille3.fr

Thierry Hamon
LIMSI-CNRS, Orsay
Université Paris 13
Sorbonne Paris Cité, France
hamon@limsi.fr

Abstract

We propose an experience on the automatic detection of sentences conveying the notion of chemical risk. Our objective is to study which resources are useful for the automatic detection of such sentences. Lexical, semantic and opinion-oriented content of the sentences is studied. Our results indicate that not only lexical and semantic content must be taken into account, but also markers related to the modality, opinion and polarity.

1 Introduction

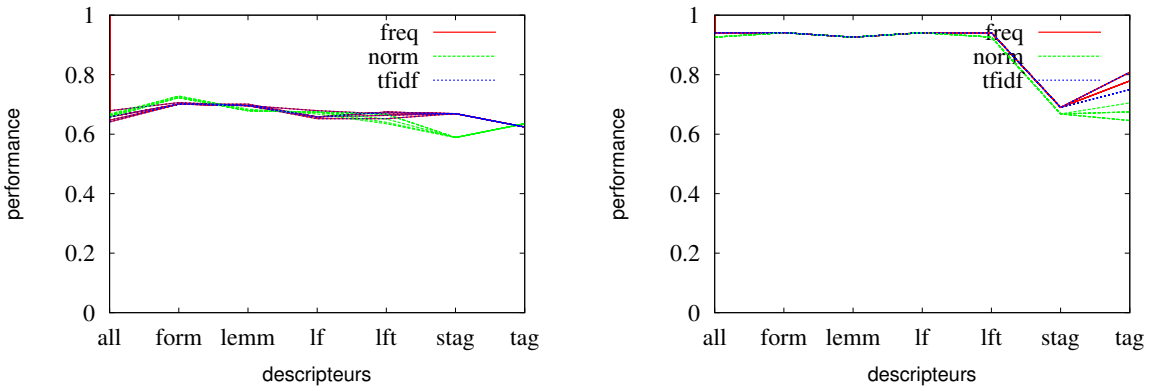
Chemical risk is relative to situations in which chemical products are dangerous for human or animal health and consumption, and for environment. The automatization of the process can help the experts to control and manage large amounts of scientific literature, that have to be analyzed to support the decision making process (van der Sluijs et al., 2008). The sentences that must be recognized are for instance: *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*, or *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*. Such sentences are to be assigned in categories related to the chemical risk: the first sentence is related to the significance of the results, while the second is related to the quality of the scientific hypothesis. If such sentences are detected in scientific publications or reports, it means that these publications or reports contain information not fully reliable and can possibly indicate the insufficiency of the corresponding studies and the presence of the risk.

The chemical risk is poorly studied, although the notion of the risk is addressed by other works: building of the dedicated resources (Makki et al., 2008), exploring of known industrial incidents (Tulechki and Tanguy, 2012), computing the exposition to the risk (Marre et al., 2010). Our objective is to study which resources are useful for the automatic detection of the sentences which convey the notion of the chemical risk.

2 Material and Methods

In addition to the lexical and semantic content of the text, we use several kinds of resources in order to favour one aspect or another. These resources contain markers oriented on modality, opinion and polarity expressed by the authors on the proposed experiements: (1) uncertainty (*possible, should, may, usually*) indicates that there are doubts on the results presented, their interpretation, etc.; (2) negation (*no, neither, lack, absent, missing*) indicates that the results have not been observed, that the study does not respect the expected norms, etc.; (3) limitations (*only, shortcoming, insufficient*) indicates that there are some limits of the work, such as unsufficient sample size, small number of tests or doses explored, etc.; (4) approximation (*approximately, commonly, estimated*) indicates other kinds of insufficiency related to imprecise values of substances, samples, dosage, etc.

The work is done with the corpus on chemical risk reporting on several chemical experiments with bisphenol A (EFSA Panel, 2010). It contains over 80,000 occurrences. The reference data are obtained through a manual categorization of the corpus sentences: 425 sentences are assigned to 55 classes of the chemical risk.



(a) Significance of the results

(b) Natural variability

Figure 1: F-measure obtained during the categorization of sentences into classes of the chemical risk.

We tackle the problem through the supervized categorization with the `Weka` platform (Witten and Frank, 2005). Sentences correspond to the units, while 7 classes (most frequent) of the chemical risk are the categories to which the sentences have to be assigned. The resources and the linguistic annotation of corpus (Schmid, 1994) provide several descriptors. These are used to build several sets of descriptors. They represent the semantic and linguistic content of the sentences: *forms* (the forms such as they occur in the corpus), *lemmas* (lemmatized forms), *lf* (combination of forms and lemmas), *tag* (POS tags, such as nouns, verbs, adjectives), *lft* (combination of forms, lemmas and POS-tags), *stag* (semantic tags of words, such as uncertainty, negation, limitations), *all* (combination of all the descriptors available). The descriptors are weighted with various methods (*freq* raw frequency, *norm* normalization by the length of the sentences, and *tfidf* tf-idf normalization).

3 Results

Figure 1 presents some results obtained for two categories: *Significance of the results* and *Natural variability of the results*. We can observe some difference according to the descriptors: the exploitation of forms, semantic tags (with *Significance of the results*) and various combinations of descriptors provide results that are often better for these two categories and for other categories. We assume that these two kinds of descriptors (lexical and semantic content of corpus and the descriptors related to modality, polarity and opinion (Vinodhini and Chandrasekaran, 2012)) provide complementary views on the content and should be combined. These results also indicate that chemical risk is not fully conceptual category but is also related to subjective and contextual values.

References

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.

J Makki, AM Alquier, and V Prince. 2008. Ontology population via NLP techniques in risk management. In *Proceedings of ICSWE*.

A Marre, S Biver, M Baies, C Defreneix, and C Aventin. 2010. Gestion des risques en radiothérapie. *Radiothérapie*, 724:55–61.

H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49.

N Tulechki and L Tanguy. 2012. Effacement de dimensions de similarité textuelle pour l'exploration de collections de rapports d'incidents aéronautiques. In *TALN*, pages 439–446.

Jeroen P van der Sluijs, Arthur C Petersen, Peter H M Janssen, James S Risbey, and Jerome R Ravetz. 2008. Exploring the quality of evidence for complex and contested policy decisions. *Environ. Res. Lett.*, 3(2).

G Vinodhini and RM Chandrasekaran. 2012. Sentiment analysis and opinion mining: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6):282–292.

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