Towards an automatic detection of the chemical risk statements

Natalia Grabar[1], **Laura Maxim**[2], **Thierry Hamon**[3]

(1) CNRS UMR 8163 STL, Université Lille 1&3, France
(2) Institut des Sciences de la Communication, CNRS UPS 3088, France
(3) LIM&BIO (EA3969), Université Paris 13, Sorbonne Paris Cité, France
natalia.grabar@univ-lille3.fr, laura.maxim@iscc.cnrs.fr, thierry.hamon@univ-paris13.fr

Abstract

We present an experiment on the detection of the chemical risk statements in institutional documents. The method relies on linguistic annotation and exploitation of classes, which describe the risk factors, and linguistic resources (negation, limitations and uncertainty markers). The method provides promising results. It will be enriched with more sophisticated NLP processing.

1. Introduction

Early detection of chemical risks (harmful effects of chemical substances on human health or the environment), such as those related to Bisphenol A or phtalates in the scientific and institutional literature may play an important role on the decisions made on marketing of the chemical products and has important concerns to the public health and security. Given the tremendous amount of the literature to be analyzed, it becomes important to provide automatic methods for the systematic mining of this literature.

2. Material and Methods

We work with four types of material: (1) classes which describe factors related to chemical risk, (2) document to process, (3) linguistic resources, and (4) reference data. Risk factor classes describe factors like causal relationship between the chemicals and the induced risk, laboratory procedures, human factors, animals tested, exposure, etc. Each class receives a short label, such as *Form of the dose-effect relationship*, *Performance of the measurement instruments* or *Sample contamination*. The processed document has been created by EFSA (European Food Safety Authority) in 2010. It proposes a literature review on Bisphenol A-related experiments and known risks or suspicions. It contains 116 pages and over 80,000 word occurrences. This is a typical institutional report which supports the decisions for managing the chemical risk. Linguistic resources contain markers for negation (Chapman et al., 2001) (*i.e.*, *no, not, neither, lack, absent, missing*, which indicate that a result has not been observed in a study, a study did not respect the norms, etc.), uncertainty (Périnet et al., 2011) (*i.e.*, *possible, hypothetical, should, can, may, usually*, which indicate doubts about the results of a study, their interpretation, significance, etc), and limitations (*i.e.*, *only, shortcoming, small, insufficient*, which indicate limits, such as small size of a sample, small number of tests or doses, etc.). The reference data is obtained thanks to a manual annotation by a specialist of chemical risk assessment: 284 segments are extracted to illustrate 34 risk factor classes.

Figure 1 presents the main steps of the method. Preprocessing is done with the Ogmios plateform[1] and provides linguistically normalized text and class labels (tokenized, POS-tagged and lemmatized (Schmid, 1994)). Then, the text is automatically annotated with the linguistic resources. During the postprocessing, we try to make a link between class labels and the text. For this, we combine information from the annotation with linguistic resources (computed in number of the corresponding markers) and the lexical intersection between the class labels and text segments (computed in percents). For instance, in the segment: *However, no specific measures were adopted to avoid sample contamination with free BPA during analytical procedures, which therefore cannot be excluded*, we find three limitation and negation markers (*however, no, cannot*), and all the words from the class label *Sample contamination*. We test several thresholds for these two values. The final step of the method is the evaluation against the reference data.

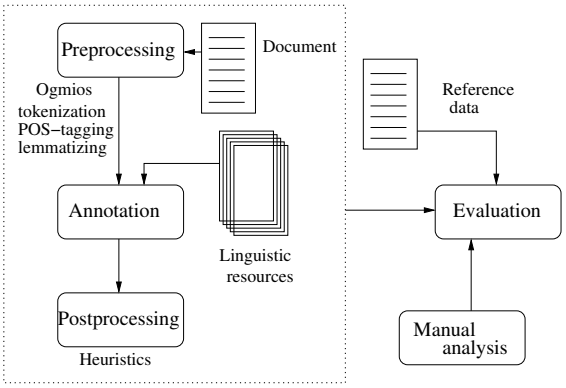


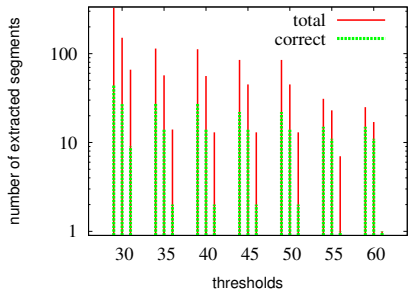
Figure 1: Main steps of the method.

3. Results and Discussion

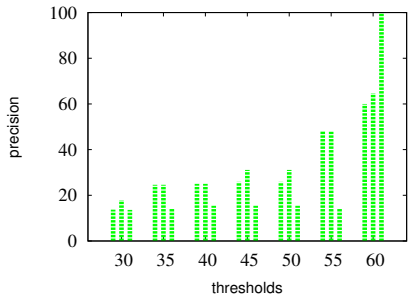
On figure 2, we present the main results obtained. On the axis x, we indicate the applied thresholds (30%, 35%, 40% etc. of words in common), while the three impulses correspond to the presence of 1, 2 or 3 markers (limitation, negation, uncertainty). With the increasing of the constraints (number of markers and percentage of common words) the number of retrieved segments dicreases while the precision increases. The best thresholds seem to be 55% or 60%: the number of segments is then important, while their precision becomes acceptable (50-65%). Here are some examples:

1. Form of the dose - effect relationship: *There was no*

[1] *http://search.cpan.org/∼thhamon/Alvis-NLPPlatform-0.6/bin/ogmios-nlp-server*



(a) Number of the extracted segments



(b) Precision of the extracted segments

Figure 2: According to the tested thresholds: number of the extracted segments and their precision.

effect on testis weight in the BPA groups, and the lack of any dose response relationship in other organ weights does not suggest a treatment-related effect

2. Choice of the experimental unit, number of animal test simultaneously: *In addition, the study has some shortcomings (small experimental groups of 3-4 animals and evaluations in males only, no indication of the number of exposed dams, or whether animals in the tested groups were littermates)*

Among the 34 classes tested, the method currently detects segments for 18 classes. We performed also a manual analysis, which showed that the method detects also segments which are correct although they are not part of the reference data. If these segments were to be considered, the precision would increase by 10 to 15%. The manual analysis revealed also the current limitations of the method. For instance, in several extracted segments, there is no syntcatic nor semantic relation between the various markers and words from labels. To mend such extractions, a syntactic analysis should be exploited. Another limitation is when there is no direct correspondence between words used in the class labels and words used in the processed document, like in *GLP compliance* and *GLP compliant*. For this a specific lexicon of synonyms and morpho-syntactic variants will be developed. Otherwise, some labels may not be evocative of their full meaning or of the expressions used in the document: other methods will be designed for them. It remains difficult to compare this experience to the existing NLP work. The closest work is done in the project Met@risk (*http://www.paris.inra.fr/metarisk*), but up to now there is no published results. Otherwise, the risk management in other domains is tackled through the building of dedicated resources (Makki et al., 2008), exploring reports on known industrial incidents and searching for similar newly created documents (Tulechki and Tanguy, 2012), calculating the exposure (Marre et al., 2010) or information extraction (Hamon and Grabar, 2010).

4. Conclusions et Perspectives

We presented results of the first experiments performed in the automatic detection of the chemical risk statements. A set of specific classes describing the factors of the chemical risk is exploited. The labels of these classes together with negation, limitation and uncertainty markers are recognized in the processed institutional document and allow to extract segments which state about the chemical danger and insufficiency of current studies. With our best thresholds, the extracted segments show precision 50-65% which may be improved if the current reference data are completed. Up to now, the method is domain independent and relies only on the labels of the classes. But this method has to evolve: new functionalities (specific contextual rules) and resources (specific synonyms and morpho-syntactic variants) will be added in order to manage more risk classes and to explore the documents more exhaustively. In order to improve the precision, we will go beyond the cooccurrences and integrate the syntactic analysis and dependencies among the words. Moreover, the method will be applied to other regulatory and scientific risk assessment reports and studies, and to other substances. The extraction results will be anaysed with several experts of the chemical risk assessment.

Acknowledgement

This work is done within the frames of the AIR REACH and DicoRisque (11-MRES-PNRPE-4-CVS-30) projects.

5. References

WW Chapman, W Bridewell, P Hanbury, GF Cooper, and BG Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform. 2001 Oct;34(5):*, 34(5):301–10.

T Hamon and N Grabar. 2010. Linguistic approach for identification of medication names and related information in clinical narratives. *J Am Med Inform Assoc*, 17(5):549–54.

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.

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

H Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *ICNMLP*, pages 44–49, Manchester, UK.

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.