Vol. 00 no. 00 2008
Pages 1–4

A New Challenge for Text Mining: Cancer Risk Assessment

Ian Lewin [1], Ilona Silins [2], Anna Korhonen [1]*, Johan Högberg [2], Ulla Stenius [2]

[1]Computer Laboratory, University of Cambridge, 15 JJ Thomson Avenue, Cambridge CB3 0FD, UK
[2]Institute of Environmental Medicine, Karolinska Institutet, S-17177, Stockholm, Sweden

ABSTRACT

Motivation: Cancer Risk Assessment (RA) of chemicals is an important and challenging multi-step task which requires combining scientific expertise with elaborate literature search and review. Due to the rapidly growing volume of RA literature, the increasing complexity of experimental evidence, and the accelerating need for chemical assessment, the task is now getting increasingly challenging to manage via manual means. Text Mining (TM) technology specifically tailored for the needs of the task could lead to considerably more systematic and efficient RA. In this paper we present the first steps taken towards the development of such technology.

Results: We have downloaded a corpus of 830 abstracts from PubMed and manually annotated the abstracts according to their relevance and the type of evidence they provide for cancer RA of selected test chemicals. The result is a taxonomy which classifies the key types of scientific evidence required for RA. The taxonomy can aid manual RA and is a starting point for the development of an approach based on TM. Using the annotated corpus we have demonstrated that supervised machine learning of large portions of the taxonomy and overall document relevance yields high accuracy and can be useful for the first step of cancer RA: finding the articles relevant for the task. We are now installing the automated classifier into the pipeline so that we can assess its impact on the RA process as a whole.

1 INTRODUCTION

The amount of scientific evidence showing a strong link between environmental chemicals and cancer calls for urgent efforts to issue exposure limits on the use of harmful chemicals. Without such precautionary actions public health and the environment are at risk. The critical tool used by authorities (e.g. governmental agencies) in making decisions on exposure limits is *Risk Assessment* (RA). Cancer RA involves examining existing published evidence to determine the relationship between exposure to a substance and the likelihood of developing cancer from that exposure (US-EPA, 2005). Performed by teams of experts in health related institutions worldwide (e.g the International Agency for Research on Cancer (IARC), the World Health Organization (WHO), the European Chemicals Agency (EHCA)), RA is a costly and challenging task which requires combining scientific expertise with elaborate literature search and review. It involves manually searching, locating and interpreting the relevant information in repositories of scientific peer reviewed journal articles - a process which can be extremely time-consuming because the data required for RA of just a single carcinogen may be scattered across thousands of journal articles. Given the exponentially growing volume of articles under inspection (e.g. PubMed expanded with over 0.5M references last year), the rapid development of molecular biology techniques, the increasing knowledge of mechanisms involved in cancer development, and the accelerating need for chemical assessment, RA is gradually getting too challenging to manage via manual means.

We are investigating a more effective approach to cancer RA based on text mining (TM). TM has been used to support various tasks in biomedicine (Ananiadou and McNaught, 2006) but to our knowledge no technology has yet been developed for cancer RA. TM could greatly assist risk assessors with the management of large textual data, increase their productivity, aid knowledge discovery, and lead into more consistent and standardized RA. From the perspective of TM, cancer RA provides a suitably complex use case for tackling the most timely problems in the field.

The task involves (1) identifying the optimal set of journal articles relevant for RA of the chemical in question and (2) studying the experimental results in these articles to determine (i) whether and (ii) exactly how the chemical causes cancer. Each step requires identifying and examining specific *types* of scientific evidence in journal articles. This is not straightforward because many articles report multiple results for several chemicals, only some of which may be relevant. Furthermore, no comprehensive and detailed specification of the range of evidence required and typically used for RA is publicly available which would enable developing a fully systematic and automatic approach.

In this first paper on the topic, we describe the work we did on identifying and organizing the key types of scientific evidence into a taxonomy. The taxonomy is based on expert annotation of a corpus of 830 abstracts downloaded from relevant PubMed journals. We also report experiments which show that the automatic classification of corpus data into classes in the taxonomy is highly accurate and can aid the first, time-consuming step of RA: the search of relevant literature.

2 CANCER RISK ASSESSMENT TAXONOMY

2.1 Data

The main types of evidence used for cancer RA are 1) scientific tests related to the carcinogenic activity: human studies (e.g. epidemiological studies), animal studies (in vivo) and cell studies (in vitro), and 2) the mode of action (MOA) of the carcinogen. The

*to whom correspondence should be addressed: alk23@cam.ac.uk

two most frequently used MOA types are *genotoxic* (i.e. chemicals cause mutations) and *non-genotoxic* (i.e. chemicals induce tumours e.g. by increasing cell proliferation).

To obtain a more comprehensive and finer-grained classification of relevant evidence, we composed a representative corpus of RA data for further analysis. Four test chemicals were first selected which are (i) well-researched using a wide range of scientific tests and (ii) represent the two most frequent MOA types: two genotoxic (1,3-butadiene, diethylnitrosamine) and two non-genotoxic (phenobarbital, diethylstilbestrol) chemicals. A set of 15 journals were then identified which are used frequently for cancer RA (e.g. Chemical Research in Toxicology, Toxicological Sciences, Mutation Research) and cover the scientific tests relevant for the task. Finally, from these 15 journals (years 1998-2008) the PubMed abstracts including the 4 test chemicals were downloaded for further analysis. We focussed on abstracts (rather than on full articles) because they are the typical starting point in RA. The resulting corpus of 830 abstracts is distributed as shown in Table 1.

2.2 Annotation Tool

An annotation tool was then designed for the analysis of the abstracts and their titles by experts in cancer RA. The tool provides two types of functionality. The first enables the experts to annotate *keywords* (words and phrases) which indicate scientific evidence relevant for examining the carcinogenic properties of chemicals. Initially a shallow taxonomy (including only the three types of tests and two types of MOA; see the above section) was integrated in the tool. Any number of keywords could be classified, and the tool also permits the same words to be classified in more than one way.

The second functionality enables to classify abstracts using the classical Information Retrieval concept of *Document Relevance*. These judgements are made at the document level. An abstract is marked as *relevant*, or *irrelevant* if the user deems after reading the title and the abstract that it is not relevant for cancer RA. Users can also mark abstracts as *unsure*. They are asked to indicate whether the relevance decision could be made simply based on the title or whether the decision required reading the abstract also.

The tool was implemented inside the Mozilla Firefox browser using its extension facility. The implementation enabled abstracts to be viewed inside a familiar web-browsing environment and to be classified by users according to their own specialized taxonomy. Previous work has observed that integrating custom functions within a familiar document browsing environment greatly encourages user uptake (Karamanis *et al.*, 2008). The RA analysed abstracts could be stored, reviewed by others and edited. In this way, the deployment of the analysis in a genuine RA scenario was able to be quickly tested.

2.3 Annotation

The annotation was carried out by three experts in cancer RA. The 830 abstracts were annotated as follows: The abstracts for two chemicals were first classified by one of the experts according to the initial shallow taxonomy and document relevance. The results were reviewed by another expert. This resulted in updates to the classification and considerable extension of the taxonomy. The entire exercise was then repeated with two further chemicals. Only very minor changes were required after this second exercise, indicating that the resulting taxonomy is relatively stable and that the agreement between the annotators is fairly good. Because it was necessary to allow discussion between the annotators in this initial work, detailed assessment of inter-annotator agreement was left for future work.

Many abstracts were classified with multiple classes in the taxonomy. One article could refer e.g. to different scientific tests as well as give MOA information. Also both MOAs (genotoxic and non-genotoxic) could occur in the same abstract (our corpus contained 17 such abstracts). For example, an abstract may refer mainly to an investigation of genotoxicity but also assess non-genotoxic modulating effects.

The experts found no difficulty in deciding on the relevance of the abstracts, in highlighting keywords, and in attaching the taxonomic concepts to pieces of text which they perceived to be relevant. When the classifications by one expert were reviewed by the other, the attachments (not just the allocated classes) proved highly valuable.

2.4 The Resulting Taxonomy and Corpus

The resulting taxonomy includes three classes at the top level. In addition to scientific tests and MOA (see section 2.1) the experts identified a third one: toxicokinetics. Each of these classes is further broken down into constituent parts. The complete taxonomy contains 45 nodes, with individual keywords falling under different nodes. The hierarchy for scientific tests is shown in Figure 1.

Table 1 shows the distribution of data across the top level of the taxonomy. Most abstracts are annotated with scientific tests (human studies being the least frequent category). Just over one-third of the abstracts are annotated for MOA, with an even distribution of genotoxic and non-genotoxic. The number of abstracts with toxicokinetic annotations is rather small. Also, at lower levels of the taxonomy, the number of data items in the more discriminating classes is fairly small. Table 2 shows the broad shape of the distribution of abstracts across leaf nodes in the taxonomy. For example, there are 6 leaf nodes under each of which more than 100 abstracts are categorized. All of these fall under scientific tests. There are also 8 leaf nodes under scientific tests with very sparse data. Although MOA abstracts are less popular overall, there are no leaf nodes with less than 20 data points. The data for toxicokinetics is small overall and nearly all located in one leaf node.

In the future, the taxonomy will be extended further by annotating data for a wider range of chemicals representing e.g. less frequently used MOA types. However, covering the main types of scientific evidence, the current taxonomy provides a good starting point for more systematic cancer RA. For example, classification of the individual articles and experimental results in the articles according to the taxonomy can be useful for risk assessors as it enables them to examine the evidence covered at any point of the workflow.

Finally, looking at the expert annotations for document relevance, just over 62% of the abstracts returned by the PubMed queries were deemed relevant for the cancer RA task by the expert reviewers (based on the title or the abstract). 10% were deemed as irrelevant and 28% were marked as unsure.

3 AUTOMATIC CLASSIFICATION EXPERIMENTS

3.1 Taxonomic Classification

To determine whether the classes in the taxonomy are machine learnable and thus optimal for TM, we trained and tested a series of

Table 1. Total of abstracts per chemical and class

Chemical	Σ	Class	Σ
1,3-butadiene	194	Evidence (total)	655
phenobarbital	270	Evidence: human	75
diethylnitrosamine	221	Evidence: animal	435
diethylstilbestrol	145	Evidence: cell	164
Total	**830**	Mode of Action (total)	287
		Genotoxic MOA	145
		Nongenotoxic MOA	159
		Toxicokinetics	56

Table 2. Abstracts per taxonomic leaf node

abstracts (f)	Σ	Evi	MoA	ToxK
$f > 100$	6	6	0	0
$50 < f \leq 100$	4	1	3	0
$20 < f \leq 50$	11	4	6	1
$f < 20$	11	8	0	3

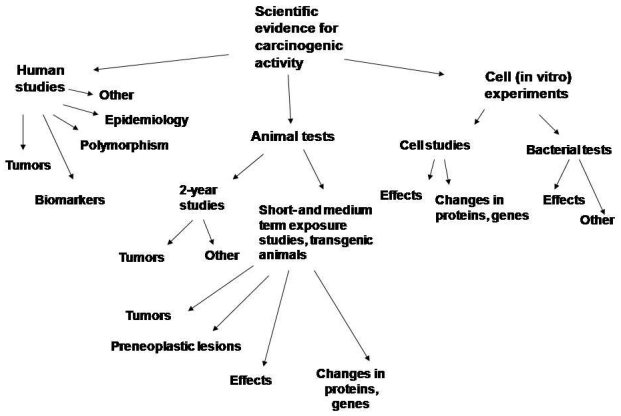

Fig. 1. Part of the Cancer Risk Assessment taxonomy

multinomial Naive Bayes classifiers on the abstracts using document level classifications. In this first experiment, we did not attempt to reproduce the classification of the text strings which serve as groundings for the classifications but only to classify abstracts as a whole. Although sophisticated classifiers, such as support vector machines, might deliver the highest accuracies attainable, Naive Bayes classifiers are optimal for our initial exploratory work because they offer good accuracy, are efficient to train, and are useful in domains in which concept drift is an issue (Manning *et al.*, 2008) (which can indeed be significant in the cancer RA domain). We used the WEKA software environment for implementation in our experiments (Witten and Frank, 2005).

Naive Bayes text classifiers aim to select the class C with maximum probability given the document d. This can be calculated via equations 1 (Bayes' rule), 2 ($Pr(d)$ is invariant across classes) and 3 (a document is considered as a bag of words and, naively, each word w is conditionally independent of the others given the class).

$$ArgMax_C\ Pr(C|d)$$

$$= \quad ArgMax_C\ \frac{Pr(C).Pr(d|C)}{Pr(d)} \tag{1}$$

$$= \quad ArgMax_C\ Pr(C).Pr(d|C) \tag{2}$$

$$= \quad ArgMax_C\ Pr(C).\prod_{w \in d} Pr(X = w|C) \tag{3}$$

$Pr(C)$ can be estimated from the frequency of documents in class C in the training corpus. In the *multinomial* model, $Pr(X = w|C)$ is estimated as the fraction of tokens in documents of class C that contain w. In addition, *add-one* smoothing is applied in the latter frequency calculation so that each word has a non-zero probability.

The words were extracted from the abstracts using very simple methods: whitespace separation for tokenization, and a standardization routine which includes lower-casing and removal of non-alphabetic, numeric or hyphen characters. We built a series of binary classifiers, one per node in the taxonomy, and performed a standard 10-fold cross-validation experiment. For each, we measured precision (P), recall (R) and F-score ($F = 2.P.R.(P+R)^{-1}$). We also experimented with an extension to multinomial Naive Bayes which has been shown to give dramatic improvements in previous work (Kibriya *et al.*, 2004). In this scheme, TFIDF scores are used in place of term frequencies in the probability estimations. $\mathrm{TFIDF}(w)$ is defined in equation 4, where w_f is the frequency of w in a document, D is the total number of documents, and d_f is the number of documents containing w.

$$TFIDFw \quad = \quad log(w_f + 1).log(\frac{D}{d_f}) \tag{4}$$

In addition, we experimented with automatic feature selection by using *Information Gain*. Information Gain measures the reduction in uncertainty about the value of the target class (the entropy) given the value of the feature selected.

3.2 Relevance

We tested whether a similar multinomial Naive Bayes classifier could prove a reliable predictor of document relevance. For this, we counted documents labelled as "unsure" simply as irrelevant and built a binary classifier. We focussed on attempting to distinguish documents that were clearly relevant because these are the documents that users are likely to look at first in the review process. We judge that it is probably more important to be reliable in directing attention to documents that deserve attention than in separating off documents that may not.

4 RESULTS

Figure 2 gives the F-scores for the top-most levels of our taxonomy for both the unextended multinomial Naive Bayes

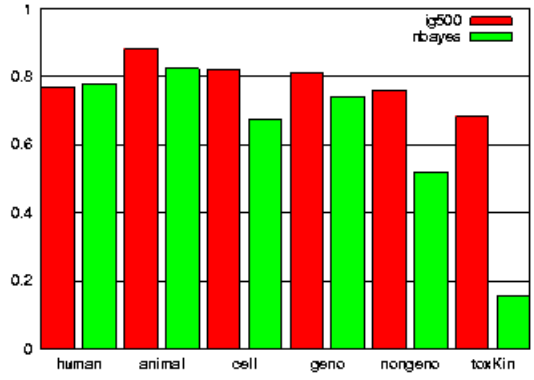

Fig. 2. F-scores per class

Table 3. Precision and recall per class

Class	F	P	R
animal	0.88	0.88	0.88
cell	0.82	0.72	0.95
human	0.77	0.66	0.92
geno	0.81	0.72	0.92
nongeno	0.76	0.68	0.86
toxK	0.68	0.61	0.77

Table 4. Mean F-score, precision and recall per taxonomic leaf node

no. of abstracts (f)	leaves	F	P	R
$f > 100$	6	0.75	0.65	0.9
$50 < f \leq 100$	4	0.79	0.68	0.95
$20 < f \leq 50$	11	0.59	0.52	0.7
$f < 20$	11	0.22	0.14	0.76

classifier (labelled *nbayes*) and the extension (*ig500*) which includes both the use of Information Gain feature selection and the use of TFIDF scores. We also experimented with the two additional features individually but reliable improvements proved possible on this dataset only with the two features operating in conjunction (data not shown). All the F-scores are promisingly above 75% with the exception of detecting documents relevant to toxicokinetics, which was also the document class for which we had the smallest training set. However, in this case the amount of improvement resulting from the multinomial Naive Bayes extensions is also most dramatic.

Table 3 breaks down the F-scores shown in Figure 2 into precision and recall. Recall is generally higher than precision and over 90% in over half the cases. The extensions to the standard classifier generally improve both measures, but recall by a greater degree.

Table 4 shows the distribution of performance measures across the leaf-nodes in the taxonomy. For example, the 6 leaf-nodes with more than 100 points each record F-scores in the vicinity of 0.75, which is the mean (the standard deviation is very low: 0.4). Good performance is also achieved for those with 50-100 data points but, as the number of data points drops below 50, performance tails off.

The document relevance classifier performed very well indeed with precision of 92.4, recall of 86.2 and F-score of 89.2

5 CONCLUSION AND FUTURE WORK

We have developed a taxonomy which classifies the key types of scientific evidence required for cancer RA. The taxonomy provides the means for more consistent manual RA and a starting point for an automatic approach based on TM. We have demonstrated that supervised machine learning of large portions of the taxonomy and overall document relevance yield good results and can be useful for the first step of cancer RA.

In the future, we plan to widen the scope of data collection beyond the four chemicals considered so far in order to extend and enrich the taxonomy further. We will compare the enriched taxonomy against the MEDLINE's Medical Subject Headings (MeSH) taxonomy to investigate the degree of overlap and the potential utility of MeSH in supporting the task. We have not yet exploited all the information made available to us by the corpus collection or extractable from the data, e.g. chemical named entities (Corbett *et al.*, 2007). Using such information we will fine-tune the classifiers to raise their performance to the best achievable. Our plan is to embed the improved classifiers into the RA workflow and evaluate, from the perspective of human-computer interaction, the impact on the work-flow and on overall task efficiency.

In the more distant future, we intend to expand on this initial work and tackle the later stages of cancer RA, for which more detailed research on RA practices and deeper linguistic analysis of the content of full journal articles will almost certainly be required.

Acknowledgements Work on this paper was funded by the Royal Society (UK), the Medical Research Council (G0601766) (UK) and the Swedish Council for Working Life and Social Research.

REFERENCES

Ananiadou, S. and McNaught, J. (2006). *Text Mining for Biology and Biomedicine*. Artech House.

Corbett, P., Batchelor, C., and Teufel, S. (2007). Annotation of chemical named entities. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 57–64, Prague, Czech Republic.

Karamanis, N., Seal, R., Lewin, I., McQuilton, P., Vlachos, A., Gasperin, C., Drysdale, R., and Briscoe, E. (2008). Natural language processing in aid of flybase curators. *BMC Bioinformatics*, **9**, 193.

Kibriya, A., Frank, E., Pfahringer, B., and Holme, G. (2004). Multinomial naive bayes for text categorization revisited. In *Australian Conference on Artificial Intelligence*, pages 488–499.

Manning, C., Raghavan, P., and Schutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. forthcoming.

US-EPA (2005). Guidelines for carcinogen risk assessment. www.epa.gov/iris/cancer032505.pdf.

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