CRAB 2.0: A text mining tool for supporting literature review in chemical cancer risk assessment

Yufan Guo[1], **Diarmuid Ó Séaghdha**[1], **Ilona Silins**[2], **Lin Sun**[1],
Johan Högberg[2], **Ulla Stenius**[2], **Anna Korhonen**[1]

[1] Computer Laboratory, University of Cambridge, UK
[2] Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

Abstract

Chemical cancer risk assessment is a literature-dependent task which could greatly benefit from text mining support. In this paper we describe CRAB – the first publicly available tool for supporting the risk assessment workflow. CRAB, currently at version 2.0, facilitates the gathering of relevant literature via PubMed queries as well as semantic classification, statistical analysis and efficient study of the literature. The tool is freely available as an in-browser application.

1 Introduction

Biomedical text mining addresses the great need to access information in the growing body of literature in biomedical sciences. Prior research has produced useful tools for supporting practical tasks such as literature curation and development of semantic databases, among others (Chapman and Cohen, 2009; Harmston et al., 2010; Simpson and Demner-Fushman, 2012; McDonald and Kelly, 2012). In this paper we describe a tool we have built to aid literature exploration for the task of chemical risk assessment (CRA). The need for assessment of chemical hazards, exposures and their corresponding health risks is growing, as many countries have tightened up their chemical safety rules. CRA work requires thorough review of available scientific data for each chemical under inspection, much of which can be found in scientific literature (EPA, 2005). Since the scientific data is highly varied and well-studied chemicals may have tens of thousands of publications (e.g. to date PubMed contains 23,665 articles mentioning phenobarbital), the task can be extremely time consuming when conducted via conventional means (Korhonen et al., 2009). As a result, there is interest among the CRA community in text mining tools that can aid and streamline the literature review process.

We have developed CRAB, an online system that supports the entire process of literature review for cancer risk assessors. It is the first and only NLP system that serves this need. CRAB contains three main components:

1. **Literature search** with PubMed integration
2. **Semantic classification** of abstracts with summary visualisation
3. **Literature browsing** with markup of information structure

These components are described further in Section 2 below. Version 2.0 of CRAB is freely available as an in-browser application; see Section 4 for access information.

2 System description

2.1 Literature search

The first step for the user is to retrieve a collection of scientific articles relevant to their need, e.g., all articles with abstracts that contain the name of a given chemical. The CRAB 2.0 search page (Figure 1) allows the user to directly query the MEDLINE database of biomedical abstracts. The search query

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations, pages 76–80, Dublin, Ireland, August 23-29 2014.

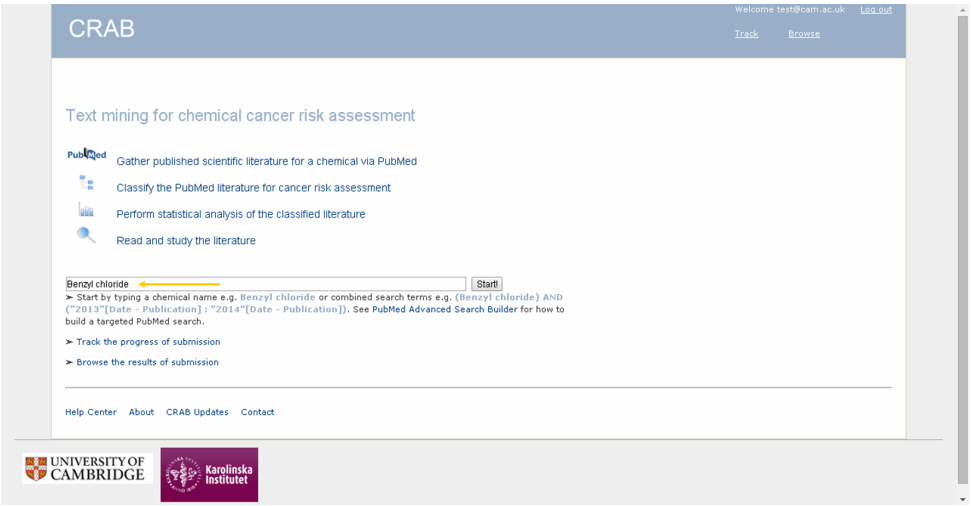
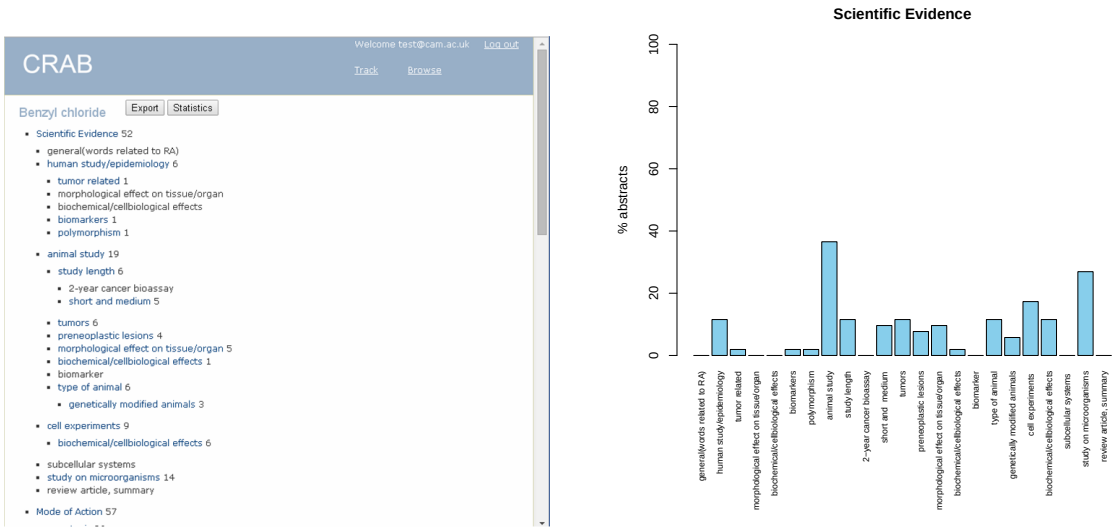


Figure 1: The CRAB 2.0 search interface



(a) Taxonomy view

(b) Histogram view

Figure 2: The CRAB 2.0 classification component

is sent, and the results received, using the E-Utilities web service provided by the National Center for Biotechnology Information.[1] This query interface supports PubMed Advanced Search, facilitating complex Boolean queries.

2.2 Semantic classification

The document collection returned by the PubMed web service is passed in XML format to a semantic classifier that annotates each abstract with 42 binary labels indicating the presence/absence of concepts relevant to CRA. These concepts are organised hierarchically in two main taxonomies: (1) kinds of scientific evidence used for CRA (e.g., *human studies*, *animal studies*, *cell experiments*, *biochemical/cell biological effects*); (2) the carcinogenic modes of action indicated by the evidence (e.g., *genotoxic*, *nongenotoxic/indirect genotoxic*, *cell death*, *inflammation*, *angiogenesis*). The underlying classifier is a support vector machine (SVM) trained on a dataset of 3,078 manually annotated abstracts. Features used by the SVM include lexical n-grams, character n-grams and MeSH concepts. For more details on the concept taxonomies, training corpus and classifier see Korhonen et al. (2012).

[1] `http://www.ncbi.nlm.nih.gov/books/NBK25501/`

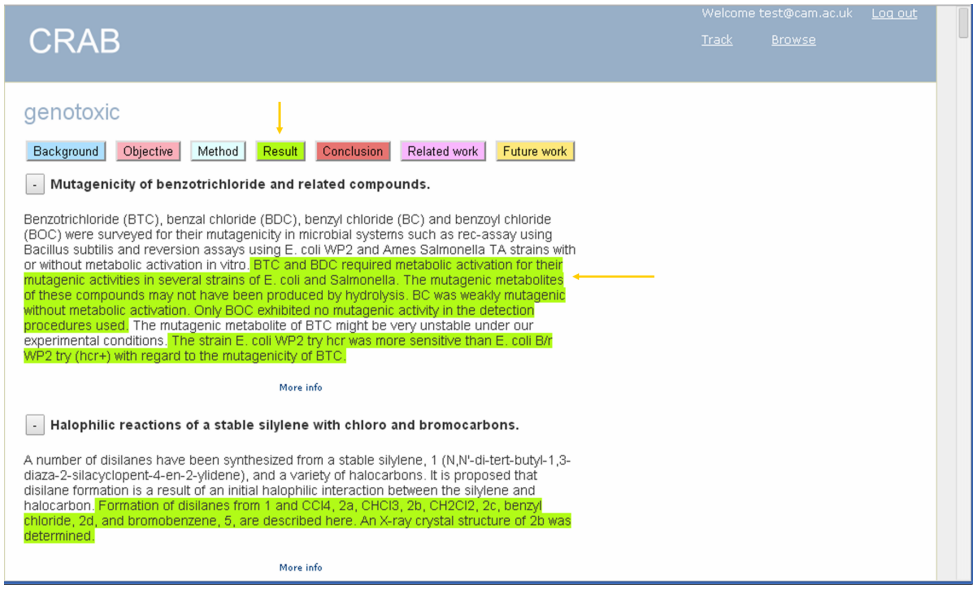


Figure 3: The CRAB 2.0 information structure component

Once each abstract in the retrieved collection has been classified, the user is presented with a summary of counts for each concept (Figure 2a). In a user study, risk assessors found this summary very useful for obtaining a broad overview of the literature, identifying groups of chemicals with similar toxicological profiles and identifying data gaps (Korhonen et al., 2012). The user can also request a histogram visualisation (Figure 2b), which is produced through a call to the statistical software R.[2]

2.3 Literature browsing

The risk assessment workflow involves close reading of relevant abstracts to identify specific information about methods, experimental details, results and conclusions. While it is not feasible to automate this process, we have shown that automatic markup and visualisation of abstracts' information structure can accelerate it considerably (Guo et al., 2011). The model of information structure incorporated in CRAB 2.0 is based on *argumentative zoning* (Teufel and Moens, 2002; Mizuta et al., 2006; Teufel, 2010), whereby the text of a scientific abstract (or article) is segmented into blocks of sentences that carry a specific rhetorical function and combine to communicate the argument the authors wish to convey to the reader. The markup scheme used in our system labels each sentence with one of seven categories: *background*, *objective*, *method*, *result*, *conclusion*, *related work* and *future work* (Guo et al., 2010). The CRAB system incorporates preprocessing (lemmatisation, POS tagging, parsing) with the C&C toolkit[3] and information structure markup with an SVM classifier that labels sentences according to a combination of lexical, syntactic and discourse features (Guo et al., 2011). The classifier has been trained on an annotated dataset of 1,000 CRA abstracts (Guo et al., 2010).

The automatic information structure markup is used to support browsing of the set of abstracts assigned a label of interest by the semantic classifier; e.g., the user can inspect all abstracts labelled *genotoxic* (Figure 3). Each information structure category is highlighted in a different colour and the user can select a single category to focus on. To our knowledge, CRAB 2.0 is the first publicly available online tool that provides information structure analysis of biomedical literature.

3 Evaluation

Intrinsic cross-validation evaluations of the semantic taxonomy classifier and information structure classifier show high performance: 0.78 macro-averaged F-score (Korhonen et al., 2012) and 0.88 accuracy (Guo et al., 2011), respectively. Furthermore, user-based evaluation in the context of real-life CRA has

[2] http://www.r-project.org/
[3] http://svn.ask.it.usyd.edu.au/trac/candc

produced promising results. (Korhonen et al., 2012) showed that the concept distributions produced by our classifier confirmed known properties of chemicals without human input. Guo et al. (2011) found that integrating information structure visualisation in abstract browsing helped risk assessors to find relevant information in abstracts 7-8% more quickly.

4 Use

CRAB 2.0 is freely available as an in-browser application at `http://omotesando-e.cl.cam.ac.uk/CRAB/request.html`. New users can register an id and password to allow them to store and retrieve data from previous sessions. Alternatively, they can use an anonymous guest account (id `guest@coling`, password `guest@coling`).

5 Conclusion

We have presented Version 2.0 of CRAB, the first NLP tool for supporting the workflow of literature review for cancer risk assessment. CRAB meets a real, specialised need and is already being used to improve the efficiency of CRA work. Although currently focused on cancer, CRAB can be easily adapted to other health risks provided with the appropriate taxonomy and annotated data for machine learning. In the future, the tool can be developed further in various ways, e.g. to support submissions in other formats than PubMed XML; to take into account journal impact factors, number of citations and cross references to better organize the literature; and to offer enriched statistical analysis of classified literature.

Acknowledgements

This work was supported by the Royal Society, Vinnova and the Swedish Research Council.

References

Wendy W. Chapman and K. Bretonnel Cohen. 2009. Current issues in biomedical text mining and natural language processing. *Journal of Biomedical Informatics*, 42(5):757–759.

EPA. 2005. Guidelines for carcinogen risk assessment. US Environmental Protection Agency.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Lin Sun, and Ulla Stenius. 2010. Identifying the information structure of scientific abstracts: An investigation of three different schemes. In *Proceedings of BioNLP-10*, Uppsala, Sweden.

Yufan Guo, Anna Korhonen, Ilona Silins, and Ulla Stenius. 2011. Weakly supervised learning of information structure of scientific abstracts: Is it accurate enough to benefit real-world tasks in biomedicine? *Bioinformatics*, 27(22):3179–3185.

Nathan Harmston, Wendy Filsell, and Michael P.H. Stumpf. 2010. What the papers say: Text mining for genomics and systems biology. *Human Genomics*, 5(1):17–29.

Anna Korhonen, Ilona Silins, Lin Sun, and Ulla Stenius. 2009. The first step in the development of text mining technology for cancer risk assessment: Identifying and organizing scientific evidence in risk assessment literature. *BMC Bioinformatics*, 10:303.

Anna Korhonen, Diarmuid Ó Séaghdha, Ilona Silins, Lin Sun, Johan Högberg, and Ulla Stenius. 2012. Text mining for literature review and knowledge discovery in cancer risk assessment and research. *PLoS ONE*, 7(4):e33427.

Diane McDonald and Ursula Kelly. 2012. The value and benefit of text mining to UK further and higher education. Report 811, JISC.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, 75(6):468–487.

Matthew S. Simpson and Dina Demner-Fushman. 2012. Biomedical text mining: A survey of recent progress. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*. Springer.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Simone Teufel. 2010. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. CSLI Publications, Stanford, CA.