CRAB Reader: A Tool for Analysis and Visualization of Argumentative Zones in Scientific Literature

Yufan Guo[1] *Ilona Silins*[2] *Roi Reichart*[1] *Anna Korhonen*[1]
(1) Computer Laboratory, University of Cambridge, UK
(2) Institute of Environmental Medicine, Karolinska Institute, Sweden
yg244@cam.ac.uk, ilona.silins@ki.se, Roi.Reichart@cl.cam.ac.uk,
alk23@cam.ac.uk

Abstract

Given the rapid publication rate in many fields of science, it is important to develop technology that can help researchers locate different types of information in scientific literature. A number of approaches have been developed for automatic identification of information structure of scientific papers. Such approaches can be useful for down-stream NLP tasks (e.g. summarization) and practical research tasks (e.g. scientific literature review), and can be realistically applied across domains when they involve light supervision. However, even light supervision requires some data annotation for new tasks. We introduce the CRAB Reader – a tool for the analysis and visualization of information structure (according to the Argumentative Zoning (AZ) scheme) in scientific literature which can facilitate efficient and user-friendly expert-annotation. We investigate and demonstrate the use of our tool for this purpose and also discuss the benefits of using the same tool to support practical tasks such as scientific literature review.

Title and Abstract in Chinese

CRAB Reader: 分析和查看科技文献论证结构的工具

随着各个学科领域出版量的迅速增长，新的文本挖掘技术的开发对研究人员从海量科技文献中寻找有用信息至为重要。许多自动识别科技文献信息结构的技术的开发，对于自然语言处理领域的下游工作（例如文本摘要） 以及实际的研究工作（例如科技文献的阅读）有很大帮助，尤其是基于弱监督学习的技术更能够实际应用于不同的学科领域。然而即使弱监督的学习依然需要一定量的人工标注数据以适应一项新任务。本文介绍了CRAB Reader： 一款分析和查看科技文献论证结构的工具。该工具有助于高效便捷地对文献进行标注。本文研究并证实了该工具对于文献信息结构的自动分析和识别有重要作用，同时讨论了该工具为实际科研工作例如科技文献的阅读带来的便利。

Keywords: Information Structure, Argumentative Zoning (AZ), Interactive Annotation, Active Learning.

Keywords in Chinese: 信息结构，论证结构分析， 交互式标注，主动学习.

Proceedings of COLING 2012: Demonstration Papers, pages 183–190,
COLING 2012, Mumbai, December 2012.

1 Introduction

There is a need to develop techniques that can help scientists locate and organize relevant information in rapidly growing scientific literature. Scientists have diverse information needs and are often interested in specific types of information in a paper. Although section headings (e.g. Methods, Results) can be an indicator of information categories of interest, many sections tend to include different types of information (e.g. the Discussion section may include information about methods and results and also provide a comparison against other peoples' work). An automatic analysis of the information category of each sentence is therefore important and can be useful for both natural language processing tasks as well as for scientists e.g. conducting literature review.

Different approaches have been developed for determining the information structure (aka. discourse, rhetorical, argumentative or conceptual structure) of scientific publications (Teufel and Moens, 2002; Mizuta et al., 2006; Shatkay et al., 2008; Teufel et al., 2009; Liakata et al., 2010; Guo et al., 2010). Some of this work has proved helpful for tasks such as information retrieval, information extraction, and summarization (Teufel and Moens, 2002; Mizuta et al., 2006; Tbahriti et al., 2006; Ruch et al., 2007). Most existing approaches are based on supervised learning and require large amounts of annotated data which limits their applicability to different domains . Recently (Guo et al., 2011b) has shown that weakly supervised learning (especially active learning) works well for determining the information structure of biomedical abstracts. This work is interesting since it can facilitate easier porting of the techniques to new tasks.

However, also approaches based on weak supervision require data annotation in real-world applications. Moreover, simulation of active learning (such as that conducted by (Guo et al., 2011b) who used a fully annotated corpus from which they restored the labels of selected sentences in each iteration) is not practical but real-time interactive annotation is needed.

This requires a development of an efficient and user-friendly annotation tool which can facilitate rapid expert-annotation of data according to categories of information structure in real-life scientific tasks. We introduce such a tool: the CRAB Reader - a tool that is capable of supporting not only off-line AZ annotation but also interactive AZ annotation through weakly supervised learning, along with visualization of information structure in scientific literature. The latter functionality can also be applied to support scientists in literature review.

The CRAB[1] reader enables analyzing biomedical articles according to the Argumentative Zoning (AZ) scheme – a scheme of information structure that describes the rhetorical progression in scientific papers (Teufel and Moens, 2002). However, since the AZ scheme has been shown to apply across different scientific domains (Teufel et al., 2009), the technology presented here can be widely applicable.

2 CRAB Reader

The CRAB Reader allows users to define a new scheme or to modify an existing scheme for information structure analysis. Currently, we use the Argumentative Zoning (AZ) scheme which was originally introduced by Teufel and Moens (2002) and which was first used to describe the rhetorical progression of scientific argument in computational linguistics papers. This scheme was subsequently adapted to other domains such as chemistry (Teufel et al., 2009) and biology (Mizuta et al., 2006). We adopt the latter version in our work, and use eight zone

[1] CRAB refers to the CRAB project which has developed text mining technology for the needs of cancer risk assessment (Korhonen et al., 2012).

categories, including Background, Problem (the research question), Method, Result, Conclusion, Connection (consistent studies), Difference (inconsistent studies) and Future-work.

Using the CRAB Reader, AZ annotation can be performed on each word, sentence, paragraph, or the entire document, depending on the requirement. Annotated papers are saved in the HTML format. Users can visualize zones in different colors in an annotated paper. CRAB Reader also supports interactive annotation which is useful for weakly supervised learning when used in conjunction with a classifier such as Support Vector Machines (SVM). More specifically, a classifier makes a request for the labels of particular sentences; in response to such a request, CRAB Reader presents the relevant sentences to the annotators, collects their annotations, and returns the labeled sentences to the classifier for further learning.

2.1 Importing articles

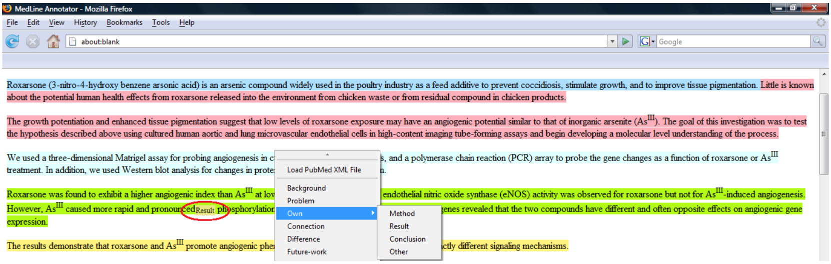
Users can import any paper in HTML format into CRAB Reader. Since the format of articles varies from journal to journal, an HTML paper needs to be transformed into XML and then formatted using XSLT (Extensible Stylesheet Language Transformations). Users need to define different style sheets for different journals. For example, the code in Figure 1 shows how to separate an abstract from the body of an article (see the <xsl:if> element), and how to format a section/subsection/paragraph (see the <xsl:for-each> element) given an article from *The Journal of Biological Chemistry*, where an <h2> tag refers to a section, an <h3/4/5> tag refers to a subsection, and a <p> tag refers to a paragraph. Currently, CRAB Reader provides style sheets for main journals on chemical risk assessment, but it can handle papers from other domains by integrating more templates. Potentially, PDF papers can also be imported into CRAB Reader after converted into HTML files, and there are various (free) tools available for converting PDF to HTML.

```
<xsl:for-each select="//div[@id='content-block']//div[@class='article fulltext-view']/div">
      <xsl:if test="contains(@class, 'section')-contains(@class, 'abstract')">
            <section>
                  <heading>
                        <xsl:value-of select="h2"/>
                  </heading>
                  <xsl:for-each select="p|h3|h4|h5">
                        <xsl:copy-of select="."/>
                  </xsl:for-each>
            </section>
      </xsl:if>
</xsl:for-each>
```

Figure 1: A fragment of the style sheet for *The Journal of Biological Chemistry*

2.2 Off-line annotation

Off-line annotation differs from interactive annotation in that users need to annotate an entire article instead of a small number of sentences selected by the machine. The off-line annotation tool in CRAB Reader is a Firefox plug-in written in XUL (XML User Interface Language). XUL is based on existing Web technologies such as CSS (Cascading Style Sheets) and JavaScript. Figure 2 shows how to use the tool in Firefox: users can select any amount of text by clicking where they want to begin, holding down the left mouse button, and then dragging the pointer over the text. Right-clicking the selected text opens a menu of zone categories such as Result, Conclusion, and so on. Users can then choose the appropriate category for annotation. The annotations are



Roxarsone (3-nitro-4-hydroxy benzene arsonic acid) is an arsenic compound widely used in the poultry industry as a feed additive to prevent coccidiosis, stimulate growth, and to improve tissue pigmentation. Little is known about the potential human health effects from roxarsone released into the environment from chicken waste or from residual compound in chicken products.

The growth potentiation and enhanced tissue pigmentation suggest that low levels of roxarsone exposure may have an angiogenic potential similar to that of inorganic arsenite (As^{III}). The goal of this investigation was to test the hypothesis described above using cultured human aortic and lung microvascular endothelial cells in high-content imaging tube-forming assays and begin developing a molecular level understanding of the process.

We used a three-dimensional Matrigel assay for probing angiogenesis in c
s, and a polymerase chain reaction (PCR) array to probe the gene changes as a function of roxarsone or As^{III}
treatment. In addition, we used Western blot analysis for changes in prote
n.

Roxarsone was found to exhibit a higher angiogenic index than As^{III} at low
endothelial nitric oxide synthase (eNOS) activity was observed for roxarsone but not for As^{III}-induced angiogenesis.
However, As^{III} caused more rapid and pronounced phosphorylation
genes revealed that the two compounds have different and often opposite effects on angiogenic gene
expression.

The results demonstrate that roxarsone and As^{III} promote angiogenic phe
ctly different signaling mechanisms.

Figure 2: Off-line annotation

saved as an attribute for each word element e.g. <w Conclusion="1">indicate</w>.

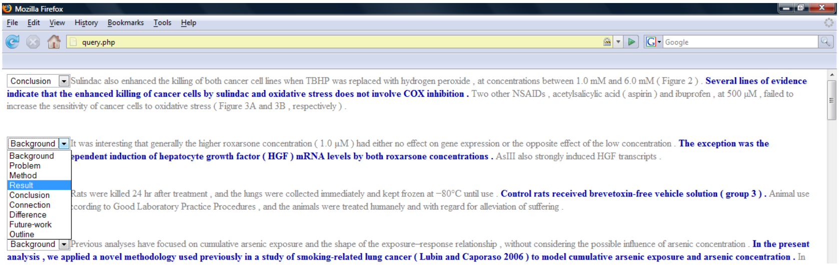
2.3 Interactive annotation via active learning

Interactive annotation is based on a client/server model, where a client (annotator) makes a request for a new assignment (unlabeled sentences); the server then trains classifiers on existing labeled data, and compares their performance on each unlabeled data, from which it selects the most informative sentences as a new assignment for the annotator. After completing the assignment the annotator submits the annotations to the server. The server then updates the training data by incorporating the fresh labels for further learning. This process can be repeated many times that is called active learning. The idea of active learning is to create a high-performance classifier but to minimize the cost of annotation.

On the server side we tested the most popular classifiers including Naive Bayes classifier, Support Vector Machines (SVM), Maximum Entropy Model, Conditional Random Fields, among many others, and SVM is so far the best classifier for this task. The features listed below were used for classification. Most of them have been successfully used in recent related work (Teufel and Moens, 2002; Mullen et al., 2005; Merity et al., 2009; Guo et al., 2011b). The C&C POS tagger and parser (Curran et al., 2007) was used for extracting syntactic features such as grammatical relations (GR) from each sentence.
Section. Normalized section names (Introduction, Methods, Results, Discussion).
Location in article/section/paragraph. Each article/section/paragraph was divided into ten equal parts. Location was defined by the parts where the sentence begins and ends.
Citation. The number of citations in a sentence (0, 1 or more).
Table and Figure. The number of referred tables and figures in a sentence (0, 1 or more).
N-gram. Any unigrams and bigrams in the corpus (an n-gram feature equals 1 if it is observed in the sentence and 0 if not; the rest of the features are defined in a similar way).
Verb. Any verbs in the corpus.
Verb Class. Verbs are grouped into 60 categories by spectral clustering (Sun and Korhonen, 2009). Each category corresponds to a feature.
Tense and Voice. Tense and voice indicated by the POS tag of main verbs and auxiliary verbs. e.g. *have|VBZ be|VBN __|VBN* indicates present perfect tense, passive voice.
Grammatical Relation. Subject (*ncsubj*), direct object (*dobj*), indirect object (*iobj*) and second object (*obj2*) relations for verbs. e.g. *(ncsubj observed difference obj)*.

Subj/Obj. The subjects/objects appearing with any verbs in the corpus.

We implemented a number of query strategies for SVM-based active learning, including least confident sampling (Lewis and Gale, 1994), margin sampling (Scheffer et al., 2001), query-by-committee (Seung et al., 1992), etc. The interactive annotation interface is a dynamic web page (HTML form) that presents a list of sentences selected by the server for human to annotate, as shown in Figure 3. It also presents the context (in gray color) of each selected sentence (in blue color) to facilitate the annotation process. After a user completes the form and clicks on the submit button, the annotations are sent to a PHP (Hypertext Preprocessor) file and then written to the server as the training data for further learning. The web page also records the annotation time for each sentence, which is a more appropriate measure of annotation effort compared to training set size. As an example, Figure 4 shows the results for real-time active learning on 50 biomedical articles. Although the curves on the left and right panels look very similar, CRAB Reader does offer an opportunity to evaluate the performance of active learning in a more natural way.



Conclusion Sulindac also enhanced the killing of both cancer cell lines when TBHP was replaced with hydrogen peroxide , at concentrations between 1.0 mM and 6.0 mM (Figure 2) . **Several lines of evidence indicate that the enhanced killing of cancer cells by sulindac and oxidative stress does not involve COX inhibition .** Two other NSAIDs , acetylsalicylic acid (aspirin) and ibuprofen , at 500 µM , failed to increase the sensitivity of cancer cells to oxidative stress (Figure 3A and 3B , respectively) .

Background It was interesting that generally the higher roxarsone concentration (1.0 µM) had either no effect on gene expression or the opposite effect of the low concentration . **The exception was the ependent induction of hepatocyte growth factor (HGF) mRNA levels by both roxarsone concentrations .** AsIII also strongly induced HGF transcripts .

Rats were killed 24 hr after treatment , and the lungs were collected immediately and kept frozen at −80°C until use . **Control rats received brevetoxin-free vehicle solution (group 3) .** Animal use cording to Good Laboratory Practice Procedures , and the animals were treated humanely and with regard for alleviation of suffering .

Background Previous analyses have focused on cumulative arsenic exposure and the shape of the exposure–response relationship , without considering the possible influence of arsenic concentration . **In the present analysis , we applied a novel methodology used previously in a study of smoking-related lung cancer (Lubin and Caporaso 2006) to model cumulative arsenic exposure and arsenic concentration .** In

Figure 3: Interactive annotation

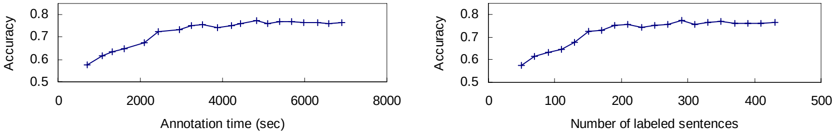
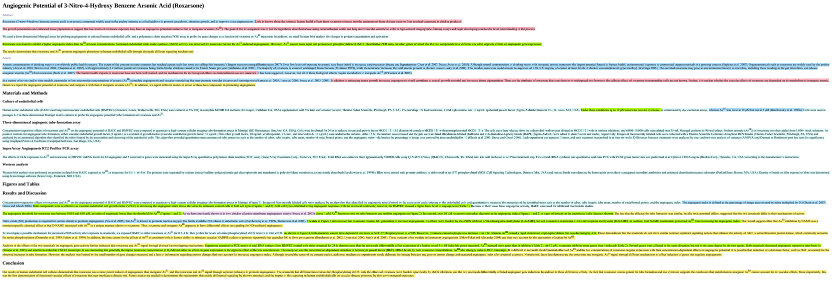


Figure 4: Performance of interactive annotation and active learning

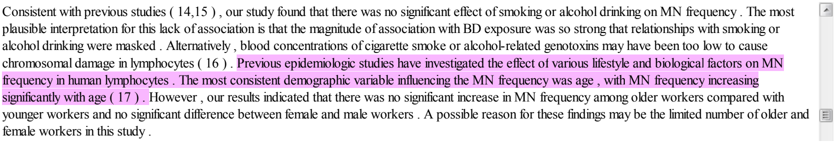
2.4 Visualizing argumentative zones

We use CSS to specify the look and format of an annotated paper. Zones have different colors as shown in Figure 5. The zone label of any particular word will come up when you mouse over it, as shown in Figure 2. The advantage of using CSS is that it provides more flexibility for the visualization of zones and allows the same annotated paper to be presented in different styles for different users (see Figure 6). Also, since CSS separates the format from the content of a document, it makes it easier to insert automatic annotations into plain text so that we can visualize the information structure of any paper on demand.



Angiogenic Potential of 3-Nitro-4-Hydroxy Benzene Arsonic Acid (Roxarsone)
Abstract
Article
Materials and Methods
Culture of endothelial cells
Three-dimensional angiogenic tube-formation assay
SuperArray Angiogenesis RT2 Profiler PCR array
Western analysis
Figures and Tables
Results and Discussion
Conclusion

Figure 5: Visualization of all zones



Consistent with previous studies (14,15) , our study found that there was no significant effect of smoking or alcohol drinking on MN frequency . The most plausible interpretation for this lack of association is that the magnitude of association with BD exposure was so strong that relationships with smoking or alcohol drinking were masked . Alternatively , blood concentrations of cigarette smoke or alcohol-related genotoxins may have been too low to cause chromosomal damage in lymphocytes (16) . Previous epidemiologic studies have investigated the effect of various lifestyle and biological factors on MN frequency in human lymphocytes . The most consistent demographic variable influencing the MN frequency was age , with MN frequency increasing significantly with age (17) . However , our results indicated that there was no significant increase in MN frequency among older workers compared with younger workers and no significant difference between female and male workers . A possible reason for these findings may be the limited number of older and female workers in this study .

Figure 6: Visualization of the ”Difference” zone

Conclusions and Future work

We have introduced CRAB Reader, a convenient tool for analysis and visualization of AZ for scientific literature review. Particularly, CRAB Reader supports real-time interactive annotation which makes it possible to apply weakly supervised learning to AZ for different tasks and domains. The tool has been used for creating a corpus of 50 AZ-annotated articles (8171 sentences), and has proved successful for active learning-based AZ on that corpus with 82% accuracy after labeling 500 sentences, which is just 2% lower than the accuracy of fully supervised learning.

In the future, we plan to use CRAB Reader for real-world applications of AZ such as question answering or customized summarization to speed up the literature review process. (Guo et al., 2011a) and (Guo et al., 2011c) have shown that users find the information in question from AZ-annotated abstracts significantly faster than from unannotated abstracts. However, is it realistic for full-text articles, given their high linguistic and informational complexity? We plan to conduct a similar question answering experiment to evaluate the usefulness of AZ-annotated articles in the context of a practical biomedical research task. We also plan to investigate whether AZ annotations are more informative than section headings for creating customized summaries for different research purposes.

Acknowledgments

The work reported in this paper was funded by the Royal Society (UK) and EPSRC (UK) grant EP/G051070/1. YG was funded by the Cambridge International Scholarship. IS was funded by the Swedish Governmental Agency for Innovation System.

References

Curran, J. R., Clark, S., and Bos, J. (2007). Linguistically motivated large-scale nlp with c&c and boxer. In *Proceedings of the ACL 2007 Demonstrations Session*, pages 33–36.

Guo, Y., Korhonen, A., Liakata, M., Karolinska, I. S., Sun, L., and Stenius, U. (2010). Identifying the information structure of scientific abstracts: an investigation of three different schemes. In *Proceedings of BioNLP*, pages 99–107.

Guo, Y., Korhonen, A., Liakata, M., Silins, I., Hogberg, J., and Stenius, U. (2011a). A comparison and user-based evaluation of models of textual information structure in the context of cancer risk assessment. *BMC Bioinformatics*, 69(12).

Guo, Y., Korhonen, A., and Poibeau, T. (2011b). A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 273–283.

Guo, Y., Korhonen, A., Silins, I., and Stenius, U. (2011c). Weakly supervised learning of information structure of scientific abstracts–is it accurate enough to benefit real-world tasks in biomedicine? *Bioinformatics*, 27:3179–85.

Korhonen, A., Séaghdha, D. O., Silins, I., Sun, L., Högberg, J., and Stenius., U. (2012). Text mining for literature review and knowledge discovery in cancer risk assessment and research. *PLoS ONE*, 7:e33427.

Lewis, D. D. and Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12.

Liakata, M., Teufel, S., Siddharthan, A., and Batchelor, C. (2010). Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of LREC'10*.

Merity, S., Murphy, T., and Curran, J. R. (2009). Accurate argumentative zoning with maximum entropy models. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 19–26.

Mizuta, Y., Korhonen, A., Mullen, T., and Collier, N. (2006). Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics on Natural Language Processing in Biomedicine and Its Applications*, 75(6):468–487.

Mullen, T., Mizuta, Y., and Collier, N. (2005). A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. *SIGKDD Explor. Newsl.*, 7:52–58.

Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbuhler, A., Fabry, P., Gobeill, J., Pillet, V., Rebholz-Schuhmann, D., Lovis, C., and Veuthey, A. L. (2007). Using argumentation to extract key sentences from biomedical abstracts. *Int J Med Inform*, 76(2-3):195–200.

Scheffer, T., Decomain, C., and Wrobel, S. (2001). Active hidden markov models for information extraction. In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, pages 309–318.

Seung, H. S., Opper, M., and Sompolinsky, H. (1992). Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294.

Shatkay, H., Pan, F., Rzhetsky, A., and Wilbur, W. J. (2008). Multi-dimensional classification of biomedical text: Toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, 24(18):2086–2093.

Sun, L. and Korhonen, A. (2009). Improving verb clustering with automatically acquired selectional preference. In *Proceedings of EMNLP*, pages 638–647.

Tbahriti, I., Chichester, C., Lisacek, F., and Ruch, P. (2006). Using argumentation to retrieve articles with similar citations. *Int J Med Inform*, 75(6):488–495.

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

Teufel, S., Siddharthan, A., and Batchelor, C. (2009). Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of EMNLP*, pages 1493–1502.