Towards an Academic Abstract Sentence Classification System

Connor Stead[1](✉), Stephen Smith[1], Peter Busch[1], and Savanid Vatanasakdakul[2]

[1] Macquarie University, Sydney, NSW, Australia
connor.stead@hdr.mq.edu.au,
{stephen.smith,peter.busch}@mq.edu.au
[2] Carnegie Mellon University, Doha, Qatar
savanid@cmu.edu

Abstract. This research in progress paper introduces a novel academic abstract sentence classification system intended to improve researcher literature discovery efficiency. The system provides three key functions: 1) displays abstracts with visual identification of each sentence's indicated literature characteristic class, 2) conversion of unstructured abstracts into structured variants and 3) categorised class sentence extraction available for export to CSV alongside literature metadata. This functionality is made possible by a web application connected to a Python instance via PHP, integration with an open access literature index via an API and a deployed academic abstract sentence classification model. The contribution of the proposed system is its ability to enhance researcher literature discovery. This paper provides context and motivation behind the development of the system, outlines its functionality and provides an outlook for future research.

Keywords: Abstracts · Literature discovery · Abstract sentence classification

1 Introduction

As a result of the large volume of academic literature available on the Internet [9], identifying literature relevant to a research undertaking can be a tedious task. This is due to the information overload associated with unprecedented widespread accessibility to literature. Even though academic literature indices and databases provide access to a significant number of digitally accessible literature, junior researchers are often at a loss as to where to begin searching for content and experienced academics often find themselves in echo chambers seeking an alternative method to identify novel research.

This paper introduces a system which can assist researchers hone in on literature within their research scope. This is achieved through the novel deployment of academic abstract sentence classification modelling into a software system designed specifically for researchers. Such modelling is an artefact of the computer science research field of natural language programming, concerned with the classification of academic abstract sentences into structured abstract format classes. Examples of these classes include 'Purpose', 'Methodology', 'Findings' and 'Contributions'. The deployment of this modelling

F. Dalpiaz et al. (Eds.): RCIS 2020, LNBIP 385, pp. 562–568, 2020.
https://doi.org/10.1007/978-3-030-50316-1_39

capability enables the software to provide three primary functions: 1) display abstracts with visual identification of each sentence's indicated literature characteristic class, 2) conversion of unstructured abstracts into structured variants and 3) categorised class sentence extraction available for export to CSV alongside literature metadata. These functions are intended to enable researchers to utilize the advancements in academic abstract sentence classification modelling to enhance literature discovery capability and improve literature review efficiency. The system proposed is the first known example of the deployment of academic abstract sentence classification capability into a software system specifically for academic researchers and with demonstrated integration with an academic literature index.

This paper is structured as follows. Firstly, some context on structured abstracts and academic abstract sentence classification will be outlined as well as the motivations for the development of the system. The paper will then introduce the system, before outlining our ongoing research on its development and the study of its utilisation.

2 Background

This section will introduce structured abstracts and academic abstract sentence classification modelling, key concepts underpinning the utility of the proposed system.

The approach to authoring abstracts is not universal across academic disciplines. Some journals and conferences enforce a structured approach, requiring a set of literature characteristics to be explicitly identified. The set of characteristics utilised in a structured abstract is known as a format [13]. Common examples of structured abstract formats from the biomedical discipline include IMRAD (‘Introduction’, ‘Methods’, ‘Results’ and ‘Discussion’) and the 8-heading format, which varies from IMRAD with the inclusion of ‘Setting’, ‘Patients’, ‘Interventions’ and ‘Outcome Measurements’.

The utility of structured abstracts has been well documented in the literature. Of particular note is the understanding that the adoption of structured abstracts increases relevant literature discovery capability [1, 7, 10, 11]. For example, Budgen et al. [2] conducted quantitative research on the utility of structured abstracts through a survey of 64 researchers and students. They determined that non-structured free text abstracts “are likely to omit substantial amounts of relevant information” (p. 457) and that structured variants “are significantly more complete and clearer than unstructured abstracts” (p. 457). These findings are supported by further research conducted by Budgen et al. [3].

The adoption of structured abstracts also benefits researchers in the computer science research field of natural language processing, as they provide a unique source of categorised sentence level observations suitable for training machine/deep learning models. These models are capable of classifying non-structured abstract sentences into structured heading format classes, such as ‘Purpose’, ‘Method’, ‘Findings’ and ‘Contributions’. Table 1 outlines state-of-the-art academic abstract sentence classification models identifying the origin paper , algorithm/modelling approach adopted and some performance characteristics. The models shown demonstrate the high-performance capability of these artefacts, most of which reach ~90% precision when classifying biomedical structured abstract sentences sourced from the PubMed 20k/200k [5] datasets.

Table 1. Summary of state-of-the-art academic abstract sentence classification models

Paper	Algorithm	Performance
Dernoncourt et al. [6]	Neural Network	PubMed 200k [5]: F1-score: 89.9%
Cohan et al. [4]	Bidirectional Encoder Representations from Transformers (BERT)	PubMed 20k [5]: 92.9% accuracy
Gonçalves et al. [8]	Neural Network	PubMed 20k [5]: 90.9% precision, 90.8% recall/F1-score
Jiang et al. [12]	Text convolutional neural network (CNN) + bidirectional recurrent neural network (bi-RNN)	PubMed 200k [5]: 94.4% accuracy (p. 8)

3 Motivation

Our research has not identified a system that deploys academic abstract sentence classification modelling capability, particularly for academic researchers. We also have not identified a system that demonstrates the integration of abstract sentence classification with academic literature indices or databases for on demand abstract sentence classification. Having greater access to relevant literature will ultimately save researchers' valuable time and resources. Therefore, we propose a system that can enhance the ability of researchers to find relevant literature more accurately and efficiently. We are also motivated to contribute to the information systems and computer science bodies of knowledge through the demonstration of how academic abstract sentence classification capability can be operationalised, and how it's introduction and adoption in the literature discovery activity impacts the ability of researchers to acquire relevant material in their efforts to produce novel research.

4 Proposed System

The proposed system is dependent on two key components: 1) the framework enabling the classification of abstract sentences and 2) the deployed classification model(s). The framework comprises of a web application, connected via PHP to a Python instance running Flask (http://flask.palletsprojects.com/en/1.1.x/), a web framework enabling REST request dispatching. The Python instance processes queries from users via the web application, forwarding these on to the open source academic literature index DOAJ

(Directory of Open Access Journals) to retrieve literature metadata. The connection to DOAJ serves as an example of connectivity - other literature indexes/databases may be connected in the future. Flat files (CSVs) of exported metadata from academic literature index searches can also be loaded and queried. Figure 1 provides a high level overview of the framework.

Once retrieved, abstracts are scored by the model, which in preprocessing splits each abstract's sentences before classifying them according to the programmed structured abstract headings. The model also provides confidence scores for each classification. The web application is then served enriched metadata via the Python instance, allowing three functions to be performed. Firstly, the application can visually highlight sentences according to the classification indicated by the model, as demonstrated in Fig. 2. Secondly, the application can convert unstructured abstracts into structured variants, as shown in Fig. 3. Alternatively, a CSV can be exported which contains returned literature metadata, and for each row an extract of structured heading sentences appended together. An example is shown in Fig. 4, filtered on ‘design/methodology/approach’.

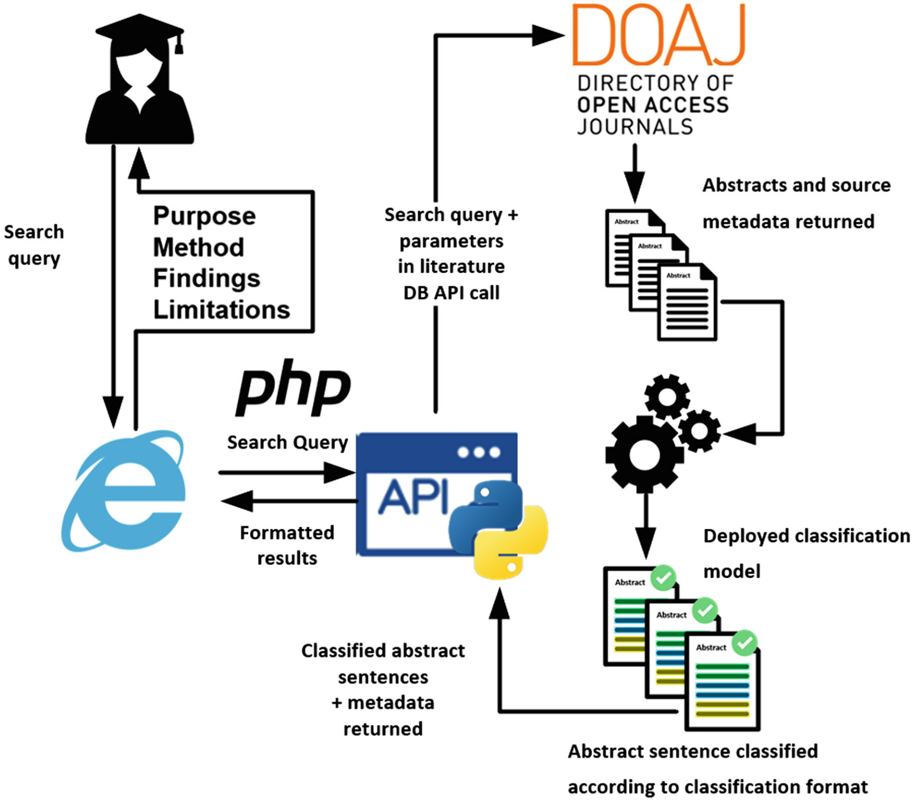
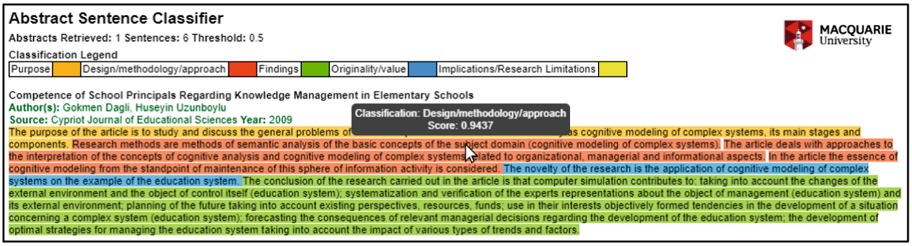

DOAJ
DIRECTORY OF OPEN ACCESS JOURNALS
Search query
Purpose
Method
Findings
Limitations
Search query + parameters in literature DB API call
Abstracts and source metadata returned
Abstract
Abstract
Abstract
php
Search Query
API
Formatted results
Deployed classification model
Abstract
Abstract
Abstract
Classified abstract sentences + metadata returned
Abstract sentence classified according to classification format

Fig. 1. High level overview of the proposed system



Abstract Sentence Classifier

Abstracts Retrieved: 1 **Sentences:** 6 **Threshold:** 0.5

Classification Legend

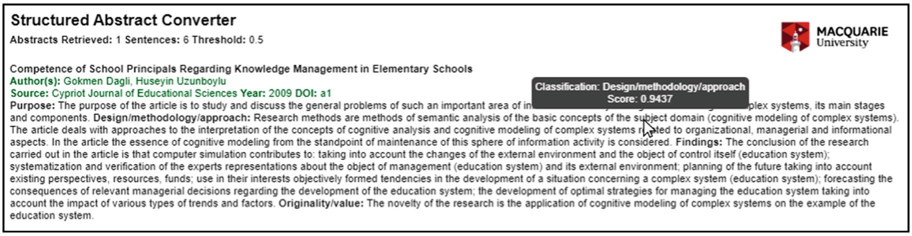
Purpose	🟧	Design/methodology/approach	🟥	Findings	🟩	Originality/value	🟦	Implications/Research Limitations	🟨

Competence of School Principals Regarding Knowledge Management in Elementary Schools
Author(s): Gokmen Dagli, Huseyin Uzunboylu
Source: Cypriot Journal of Educational Sciences **Year:** 2009

Classification: Design/methodology/approach
Score: 0.9437

The purpose of the article is to study and discuss the general problems of as cognitive modeling of complex systems, its main stages and components. Research methods are methods of semantic analysis of the basic concepts of the subject domain (cognitive modeling of complex systems). The article deals with approaches to the interpretation of the concepts of cognitive analysis and cognitive modeling of complex systems related to organizational, managerial and informational aspects. In the article the essence of cognitive modeling from the standpoint of maintenance of this sphere of information activity is considered. The novelty of the research is the application of cognitive modeling of complex systems on the example of the education system. The conclusion of the research carried out in the article is that computer simulation contributes to: taking into account the changes of the external environment and the object of control itself (education system); systematization and verification of the experts representations about the object of management (education system) and its external environment; planning of the future taking into account existing perspectives, resources, funds; use in their interests objectively formed tendencies in the development of a situation concerning a complex system (education system); forecasting the consequences of relevant managerial decisions regarding the development of the education system; the development of optimal strategies for managing the education system taking into account the impact of various types of trends and factors.

Fig. 2. Highlighting of sentences according to the classified structured abstract class.



Structured Abstract Converter

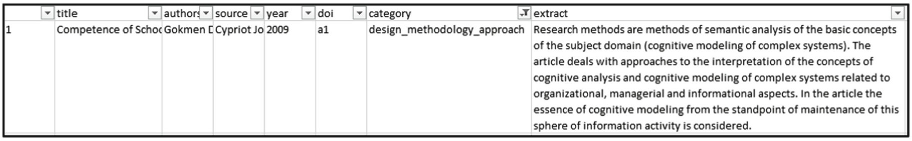
Abstracts Retrieved: 1 **Sentences:** 6 **Threshold:** 0.5

Competence of School Principals Regarding Knowledge Management in Elementary Schools
Author(s): Gokmen Dagli, Huseyin Uzunboylu
Source: Cypriot Journal of Educational Sciences **Year:** 2009 **DOI:** a1

Classification: Design/methodology/approach
Score: 0.9437

Purpose: The purpose of the article is to study and discuss the general problems of such an important area of in plex systems, its main stages and components. **Design/methodology/approach:** Research methods are methods of semantic analysis of the basic concepts of the subject domain (cognitive modeling of complex systems). The article deals with approaches to the interpretation of the concepts of cognitive analysis and cognitive modeling of complex systems related to organizational, managerial and informational aspects. In the article the essence of cognitive modeling from the standpoint of maintenance of this sphere of information activity is considered. **Findings:** The conclusion of the research carried out in the article is that computer simulation contributes to: taking into account the changes of the external environment and the object of control itself (education system); systematization and verification of the experts representations about the object of management (education system) and its external environment; planning of the future taking into account existing perspectives, resources, funds; use in their interests objectively formed tendencies in the development of a situation concerning a complex system (education system); forecasting the consequences of relevant managerial decisions regarding the development of the education system; the development of optimal strategies for managing the education system taking into account the impact of various types of trends and factors. **Originality/value:** The novelty of the research is the application of cognitive modeling of complex systems on the example of the education system.

Fig. 3. Structuring of an unstructured abstract. Heading classes are identified in bold.



	title	authors	source	year	doi	category	extract
1	Competence of Schoo	Gokmen D	Cypriot Jo	2009	a1	design_methodology_approach	Research methods are methods of semantic analysis of the basic concepts of the subject domain (cognitive modeling of complex systems). The article deals with approaches to the interpretation of the concepts of cognitive analysis and cognitive modeling of complex systems related to organizational, managerial and informational aspects. In the article the essence of cognitive modeling from the standpoint of maintenance of this sphere of information activity is considered.

Fig. 4. Structured class content and metadata extracted to CSV. Filtered on 'findings' extracts.

To demonstrate the utility of the system we trained a sentence classification model using the XLNet [15] modelling method, which has achieved state-of-the-art status for several classification challenges [15]. A XLNet model was trained using the Emerald 20k dataset [14], containing 201,452 classified sentences from 20,000 multidisciplinary abstracts. Sentences in the dataset are classified into the following structured abstract heading classes: purpose, design/methodology/approach, originality/value, practical implications, social implications and research limitations/implications. The model achieved 73.3% precision when tested on a holdout subset. Whilst performance of the model is not state-of-the-art, it served the purpose of demonstrating the system's ability to classify academic abstract sentences on demand. Future system development will look towards the deployment of state-of-the-art models, such as those identified in Table 1. This highlights the adaptability of the system, in that it permits the deployment of pre-trained models that enable classification via a Python instance. It is also possible to deploy several classification models and for alternate models to be used for scoring depending on characteristics of the abstract or query, such as the origin discipline.

5 Ongoing Research

We are conducting research exploring the utility of the system, specifically what role it plays as a facilitating tool in the literature discovery activity conducted by multidisciplinary researchers. Our examination will adopt a theoretical framework which permits the analysis of researchers adopting the system into their research efforts. We anticipate this research will yield both quantitative and qualitative insights into the value of the academic abstract sentence classification, thereby directing any future research and development into the deployment of such capability. We will also be working to enhance the system through state-of-the-art model deployment, user experience improvement as well as alternative academic literature index and database integration.

Acknowledgments. The authors wish to acknowledge the Australian Government Research Training Program Scholarship which enabled this research to take place.

References

1. Bayley, L., Eldredge, J.: The structured abstract: an essential tool for researchers. Hypothesis **17**(1), 11–13 (2003). PMID: 15858627
2. Budgen, D., Kitchenham, B.A., Charters, S.M., Turner, M., Brereton, P., Linkman, S.G.: Presenting software engineering results using structured abstracts: a randomised experiment. Empirical Softw. Eng. **13**(4), 435–468 (2008). https://doi.org/10.1007/s10664-008-9075-7
3. Budgen, D., Burn, A.J., Kitchenham, B.: Reporting computing projects through structured abstracts: a quasi-experiment. Empirical Softw. Eng. **16**(2), 244–277 (2011). https://doi.org/10.1007/s10664-010-9139-3
4. Cohan, A., Beltagy, I., King, D., Dalvi, B., Weld, D.S.: Pretrained language models for sequential sentence classification (2019). arXiv preprint arXiv:1909.04054
5. Dernoncourt, F., Lee, J.Y.: PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts (2017). arXiv preprint arXiv:1710.06071
6. Dernoncourt, F., Lee, J.Y., Szolovits, P.: Neural networks for joint sentence classification in medical paper abstracts (2016). arXiv preprint arXiv:1612.05251
7. Eldredge, J.: Evidence-based librarianship: the EBL process. Libr. Hi Tech **24**(3), 341–354 (2006). https://doi.org/10.1108/07378830610692118
8. Gonçalves, S., Cortez, P., Moro, S.: A deep learning classifier for sentence classification in biomedical and computer science abstracts. Neural Comput. Appl. **32**, 6793–6807 (2019). https://doi.org/10.1007/s00521-019-04334-2
9. Gusenbauer, M.: Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics **118**(1), 177–214 (2019). https://doi.org/10.1007/s11192-018-2958-5
10. Hartley, J.: Is it appropriate to use structured abstracts in social science journals? Learn. Publish. **10**(4), 313–317 (1997). https://doi.org/10.1087/09531519750146789
11. Hartley, J., Sydes, M., Blurton, A.: Obtaining information accurately and quickly: are structured abstracts more efficient? J. Inform. Sci. **22**(5), 349–356 (1996). https://doi.org/10.1177/016555159602200503
12. Jiang, X., Zhang, B., Ye, Y., Liu, Z.: A hierarchical model with recurrent convolutional neural networks for sequential sentence classification. In: Tang, J., Kan, M.-Y., Zhao, D., Li, S., Zan, H. (eds.) NLPCC 2019. LNCS (LNAI), vol. 11839, pp. 78–89. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32236-6_7

13. Nakayama, T., Hirai, N., Yamazaki, S., Naito, M.: Adoption of structured abstracts by general medical journals and format for a structured abstract. J. Med. Libr. Assoc. **93**(2), 237 (2005). PMID: 15858627
14. Stead, C., Smith, S., Busch, P., Vatanasakdakul, S.: Emerald 110k: a multidisciplinary dataset for abstract sentence classification. Paper presented at the Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association, pp. 120–125 (2019)
15. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: generalized autoregressive pretraining for language understanding. In: 33rd Conference on Neural Information Processing Systems, pp. 5754–5764 (2019)