Partridge: An Effective System for the Automatic Cassification of the Types of Academic Papers

James Ravenscroft, Maria Liakata and Amanda Clare

Abstract Partridge is a system that enables intelligent search for academic papers by allowing users to query terms within sentences designating a particular core scientific concept (e.g. Hypothesis, Result, etc). The system also automatically classifies papers according to article types (e.g. Review, Case Study). Here, we focus on the latter aspect of the system. For each paper, Partridge automatically extracts the full paper content from PDF files, converts it to XML, determines sentence boundaries, automatically labels the sentences with core scientific concepts, and then uses a random forest model to classify the paper type. We show that the type of a paper can be reliably predicted by a model which analyses the distribution of core scientific concepts within the sentences of the paper. We discuss the appropriateness of many of the existing paper types used by major journals, and their corresponding distributions. Partridge is online and available for use, includes a browser-friendly bookmarklet for new paper submission, and demonstrates a range of possibilities for more intelligent search in the scientific literature. The Partridge instance and further information about the project can be found at http://papro.org.uk.

1 Introduction

Since the advent of the 'Digital Age', the amount of information available to researchers has been increasing drastically, relevant material is becoming progressively more difficult to find manually and the need for an automated information

J. Ravenscroft (✉) · A. Clare
Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, UK
e-mail: ravenscroft@papro.org.uk

A. Clare
e-mail: afc@aber.ac.uk

M. Liakata
Department of Computer Science, University of Warwick, Aberystwyth, UK
e-mail: M.Liakata@warwick.ac.uk

M. Bramer and M. Petridis (eds.), *Research and Development in Intelligent Systems XXX*,
DOI: 10.1007/978-3-319-02621-3_26, © Springer International Publishing Switzerland 2013

retrieval tool more apparent. There are already a large number of information retrieval and recommendation systems for scientific publications. Many of these systems, such as AGRICOLA,[1] the Cochrane Library[2] and Textpresso[3] index only publications from predefined journals or topics (for the above examples, Agriculture, Biology and Bioinformatics respectively). Unfortunately, these domain specific indexing systems usually only contain a small subset of papers, excluding potentially crucial literature because it does not quite fit into the subject domain.

The value of these systems to their users is often restricted by the small proportion of available literature that they index, forcing researchers to use multiple, domain specific, search engines for their queries. In contrast, there are also a number of interdisciplinary indexing systems and online journals such as arXiv[4] and PloSOne[5] that try to incorporate wide ranges of papers from as many disciplines as possible. The traits of these systems often complement those of their domain-specific counterparts; they provide a comprehensive collection of literature but insufficient filtering and indexing capabilities, usually based on title and abstract.

However, the document title is just one of the crucial parts of a scientific paper's structure. Liakata et al. describe a system for automatically processing and classifying sentences in a research paper according to the core scientific concept (CoreSC) that they describe [1]. There are 11 CoreSCs, including *Hypothesis*, *Goal*, *Background*, *Method*, *Result* and *Conclusion*. CoreSC labels can be allocated to all sentences in a scientific paper in order to identify which scientific concept each sentence encapsulates. SAPIENTA[6] is a publicly available machine learning application which can automatically annotate all sentences in a scientific paper with their CoreSC labels. It was trained using a corpus of physical chemistry and biochemistry research papers whose sentences were manually annotated using the CoreSC [2] scheme. An intelligent information retrieval system can use this data to provide better filtering and search capabilities for researchers. Partridge implements such context-aware keyword search, by allowing researchers to search for papers where a term appears in sentences with a specific CoreSC label (e.g only in *Method* sentences). This can be used to greatly improve both the precision with which researchers are able to perform searches for scientific literature and the accuracy of those searches in terms of relevance to the reader.

The type of a paper (*Review*, *Case Study*, *Research*, *Perspective*, etc) is another useful feature through which a user can narrow down the results of a search. The type of a paper can then be used to augment queries. For example, a user may search for a *Review* paper containing the keywords “PCR microfluidics”, or a *Research* paper with a *Hypothesis* containing the keywords “cerevisiae” and “glucose”. Such paper types are not yet standardised by journals. We expect the structure of a paper

[1] http://agricola.nal.usda.gov/

[2] http://www.thecochranelibrary.com/

[3] http://www.textpresso.org/

[4] http://arxiv.org

[5] http://plosone.org

[6] http://www.sapientaproject.com

to reflect its paper type. For example, review papers would be expected to contain a large amount of background material. In this article, we describe the application of machine learning (using random forests) to create predictive models of a paper's type, using the distribution of CoreSC labels found in the full text of the paper.

This model of paper type is currently in use in our Partridge system, which has been created as an intelligent full-text search platform for scientific papers. Partridge (which currently holds 1288 papers and is constantly expanding) makes use of automatically derived CoreSC sentence labels and automatically derived paper types, to allow deeper information queries. We discuss the reliability of this model of paper types and the insights that have been gained for the authorship of papers.

2 Methods

2.1 Collection of Scientific Articles

Partridge allows users to upload any paper which is free of copyright restrictions. For the purpose of this study we needed a large set of papers that we could label with CoreSC to investigate how this information assists classification into paper type. Open Access (OA) journals provide free read access to their papers, but many do not permit the user of articles for data mining purposes. The Public Library of Science (PLoS) journals contain large volumes of OA literature under a permissive license that allows data mining. They also use the PubMed Central markup schema, which is compatible with SAPIENTA, for papers published through their journal. The PLoS journals advanced search offers approximately 50 types of paper through which to restrict the search. Many of these paper type categories contain too few papers to be useful, others are too ad hoc (e.g. *Message from PLoS*). Indeed it is not clear how these paper types have been identified. We chose to look at a range of types, some of which we expect to overlap or have an unclear distinction. These are namely: Essay, Correspondence, Case Study, Perspective, Viewpoint, Opinion, Review, Research.

A script called *plosget.py*[7] was written, in order to download the papers via the PloS RESTful search API.[8] The number of papers downloaded per paper type category was as follows: 200 Essay, 99 Correspondence, 107 Case Study, 200 Perspective, 74 Viewpoint, 93 Opinion, 312 Review, 200 Research. These formed a corpus of 1285 check the numbers add up papers.

Figure 1 shows the CoreSC content of a review paper and a research paper randomly selected from the corpus. The review papers tend to be made up almost entirely from Background CoreSC sentences. However, research papers are much more evenly spread, made up of several different types of CoreSC. This

[7] https://github.com/ravenscroftj/partridge

[8] http://api.plos.org/solr/faq/

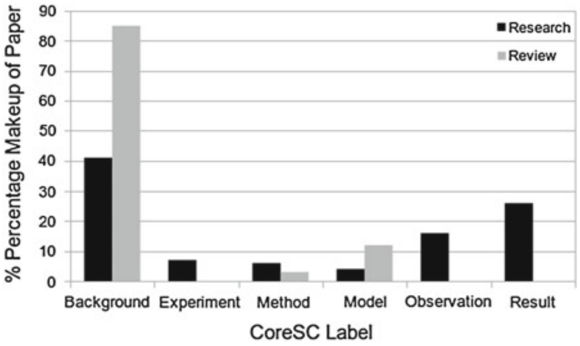

Fig. 1 The CoreSC makeup of a research paper and a review paper randomly selected from the corpus

investigation suggested that there is almost certainly a discriminative relationship between CoreSC categories and a paper's type.

2.2 Paper Processing and Classification

In order to obtain automatic labeling of sentences in a paper with CoreSC concepts we first needed to convert the papers to a format that the the CoreSC classifier, SAPIENTA can analyse. Currently, SAPIENTA supports the SciXML and PubMed Central DTDs. Papers in PDF format are first converted to XML using PDFX, a free service hosted by the University of Manchester [3]. Once the paper has been split into sentences with Liakata's SSplit tool [4], SAPIENTA is used to annotate each sentence with a relevant CoreSC label. These labels are assigned using a conditional random fields model based upon a number of features such as the sentence's location within the paper and pairs and triplets of words found consecutively within the sentence [1]. The SAPIENTA system has been trained on a corpus of 265 chemistry and biochemistry papers and we have done no domain adaptation for the papers in PLoS.

Random forest learning [5] was chosen for the construction of the paper type classifier. This is because random forests are a fast, accurate and widely accepted learning technology, and the underlying decision trees can be inspected to understand some of the reasoning behind the predictions made by the final paper type classifiers. The feature set consisted of the percentage composition of each of the 11 respective CoreSC labels assigned to the sentences in each paper i.e. (the percentage of *Background* sentences, the percentage of *Hypothesis* sentences, etc.). The random forest learning was conducted using the Orange data mining library for Python [6]. The parameters for the random forest learner were as the defaults, except with min_subsets = 5 and same_majority_pruning = true. We used 10-fold cross validation to estimate the precision, recall and F-measure.

3 Results and Discussion

The results of the random forest learning as recall, precision and F-measure for each paper type, averaged over a 10-fold cross validation are shown in Table 1. Paper types *Research*, *Review* and *Correspondence* are the most accurate classes, and *Viewpoint* is the most difficult paper type to predict.

A confusion matrix for the paper types is given in Table 2. Research papers are usually predicted to belong to the Research class, and are sometimes confused with Case Study. However, Case Study papers are often predicted to be Research or Review. The ratio between Case Study, Research and Review papers is 1:2:3 so the outcome is not entirely surprising in terms of the data size effect. We expected Research and Case Study to share similar paper structures but perhaps the fact that Case Study is equally confused with Research and Review suggests two distinct types of Case Study. We expected the classes Essay, Opinion, Perspective and Viewpoint to be confused, as the four labels all indicate a paper containing an author's personal thoughts on an issue, rather than experiment-driven science. Opinion is confused with Perspective and Viewpoint. Perspective is confused with almost all classes, except Research. Viewpoint is a small class and mostly confused with Opinion and Perspective. From this it would seem that the paper types {Opinion, Perspective, Viewpoint} are very similar and should be grouped together into one category. Indeed when we trained the model on a super class consisting of the previous three categories, performance was improved overall with the following F-measures: OpinionSuper: 59, Research: 67.6, Review: 63.7, Essay: 44.9, Correspondence: 46 and Case Study 24.8.

We inspected the detail of a single decision tree, grown on the entire dataset, to gain further insight into which CoreSC class decisions were responsible for the paper

Table 1 Per-class recall, precision and F-measure micro-averaged over 10-fold cross validation, reporting the results on the held-out validation segment

Classes	Recall (%)	Precision (%)	F-measure (%)
Case study	24.3	24.5	24.4
Correspondence	**54.5**	**50.5**	**52.4**
Essay	45.0	46.2	45.6
Opinion	24.7	23.5	24.1
Perspective	25.5	30.4	27.7
Research	**70.0**	**61.9**	**65.7**
Review	**63.1**	**62.5**	**62.8**
Viewpoint	14.9	15.7	15.3

Top paper types in bold

Table 2 Confusion matrix summed over 10-fold cross validation, reporting the results on the held-out validation segment

	Case study	Correspondence	Essay	Opinion	Perspective	Research	Review	Viewpoint
Case study	26	2	11	2	5	26	25	10
Correspondence	4	54	3	9	15	8	4	2
Essay	8	5	90	2	39	5	46	5
Opinion	4	7	2	23	27	10	3	17
Perspective	11	16	49	30	51	6	26	11
Research	19	6	2	8	5	140	12	8
Review	27	8	35	5	13	21	197	6
Viewpoint	7	9	3	19	13	10	2	11

Rows represent true classes, columns represent predicted classes type predictions. This was a very large tree, with a depth of 37 nodes in places. The first decision was based upon the number of Background sentences in the article. Low Background percentages, of less than 0.694 indicates a Correspondence paper. 28 out of the 31 Correspondence papers were correctly classified by this decision.

The next decision, for higher amounts of Background, was based on the percentage of Experiment sentences in the paper. For a very low percentage of Experiment sentences (<0.061), the papers then branched into a long side chain of detailed classification decisions, to separate mostly the Opinion, Viewpoint and Perspective papers from other Correspondence papers, and a few examples of the remaining categories. For a higher percentage of Experiment sentences, Research papers were classified as those that had Observations $>5.6\,\%$, or Conclusions $<=1.4\,\%$, whereas Case Studies had fewer Observations and more Conclusions.

The largest node classifying Essay did so via a route after the low Experiment decision that asked for a Background $>48\,\%$, but then low values for Goal, Hypothesis, Result, Observation, Model, Conclusion and Object. These decisions seem reasonable and agreed with our expectations of the content of an Essay.

4 Summary and Conclusions

To summarise, we have demonstrated that paper type can largely be predicted from the distribution of the automatically obtained CoreSC sentence labels in the full text of a paper. We have described some of the particular CoreSC features that determine a paper type (e.g. Correspondence characterised by very little Background, Opinions characterised by low percentage of Experiment) and discussed which of the paper types are not easily separable in this way. We recommended for example merging the Opinion, Viewpoint and Perspective articles into a single category, which increased overall classification performance. It is also potentially useful to distinguish between two types of Case study. Analysis of an example tree shows the decision making process to be complex, but agrees with our general understanding of paper types. Partridge allows refinement of paper search using CoreSC and paper type, both of which can be intelligently determined using machine learning methods.

The potential for automatic extraction of useful features from scientific papers to assist researchers in their knowledge queries is now an exciting area for research. Open Access journals that permit full text mining lead the way in allowing this research to expand and flourish. In future work we aim to develop and implement a range of further useful properties that will uncover more of the information that is hidden within the text of articles, and to use Partridge as a working engine to demonstrate their usefulness in practice.

Acknowledgments We thank the Leverhulme Trust for the support to Dr Liakata's Early Career Fellowship and also EMBL-EBI, Cambridge UK for the facilities offered to Dr Liakata.

References

1. M. Liakata, S. Saha, S. Dobnik, C. Batchelor, D. Rebholz-Schuhmann, Bioinformatics pp. 991–1000 (2012)
2. M. Liakata, S. Teufel, A. Siddharthan, C. Batchelor, in *Proceedings of LREC'10* (2010)
3. A. Constantin, S. Pettier, A. Voronkov, in *Proceedings of the 13th ACM Symposium on Document Engineering (Doc Eng)* (2013)
4. M. Liakata, L.N. Soldatova, et al., in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* (Association for, Computational Linguistics, 2009), pp. 193–200
5. L. Breiman, Machine, Learning pp. 5–32 (2001)
6. J. Demšar, B. Zupan, G. Leban, T. Curk, in *Knowledge Discovery in Databases PKDD 2004, Lecture Notes in Computer Science*, vol. 3202, ed. by J.F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi. Faculty of Computer and Information Science, University of Ljubljana (Springer, 2004), *Lecture Notes in Computer Science*, vol. 3202, pp. 537–539. DOI 10.1007/b100704. URL http://www.springerlink.com/index/G58613YV08BX48QJ.pdf