Can Models of Author Intention Support Quality Assessment of Content?

A J Casey[1], Bonnie Webber[1], and Dorota Głowacka[2]

[1] University of Edinburgh {a.j.casey,bonnie}@inf.ed.ac.uk
[2] University of Helsinki glowacka@cs.helsinki.fi

Abstract. Academics seek to find, understand and critically review the work of other researchers through published scientific articles. In recent years, the volume of available information has significantly increased, partly due to technological advancements and partly due to pressures on academics to 'publish or perish'. This amount of papers presents a challenge not only for the peer-review process but also for readers, particularly inexperienced readers, to find publications of high quality. Whilst one might rely on citation or journal rankings to help guide this decision, this approach may not be completely reliable due to biased peer-review processes and the fact that the citation count of an article does not *per se* indicate its quality. Here, we analyse how expected author intentions in a *Related Work* section can be used to indicate its quality. We show that author intentions can predict the quality with reasonable accuracy and propose that similar approaches could be used in other sections to provide an overall picture of quality. This approach could be useful in supporting peer-review processes and for a reader in prioritising articles to read.

Keywords: Article Quality · Author Intentions · Supporting peer-review

1 Introduction

Recent years have seen an increase in the volume of scientific publications. The amount of published material poses a challenge for the reader, in particular an inexperienced one, who must navigate this overwhelming wealth of material to find relevant and high quality content. Another challenge is for the peer-review process. There is only a limited pool of experts to undertake peer-review and the high volume of submitted material puts pressure on this limited resource. Having automated ways to assess quality could support the peer-review process and help the overwhelmed reader to prioritise their ever growing reading list.

Automating judgement of quality in research is challenging as it requires knowledge. Bridges [2] describes this judgement of research quality as a connoisseurship which draws on one's own knowledge and experience of the field. This, in turn, not only allows one to comment on specific features but also gives one the ability to appreciate the overall composition of the text. It is recognised that it would be difficult, if not impossible, to try to emulate this level of human

judgement in an automated fashion. We propose that considering how argument intentions are represented linguistically and quantifying the depth of this representation may help to build quality indicators that could prove useful in supporting the peer-review process or to help readers identify better reading material. The intuition behind using argument elements to define quality has support in existing literature with essay scores shown to be linked to argumentative elements identified through discourse analysis [4, 15].

Based on this premise, we consider *Related Work* sections from published papers as a case study. We assess these sections rating them as Good (G), Average (Avg) or Poor (P). We use *Related Work* sections annotated with author intentions designed to give content feedback [5]. We analyse the relationship of these author intentions and the quality ratings, showing that quality and author intention occurrence are related, predicting with reasonable accuracy the quality rating of a *Related Work*.

2 Related Work

Peer-review, generally accepted as the gold standard of assessing quality, is not without issue. There are problems of bias, publication delays, problems with detecting fraud and/or errors, and unethical practices [18]. Metrics, such as citations and download counts, have also been considered as indicators of quality. But these too have known issues such as dependence of the size of discipline, and they take time to accumulate. Authors and research teams have been known to carry out unnecessary self-citations to increase their own citations [8]. Despite these problems, we do not believe peer-review or using citation measures should be replaced. Rather, we see our work as an additional tool. It could, for example, be used for triage: if our tool rates a paper Poor or Good, perhaps it needs only one reviewer to confirm it, with a second one only needed if the first reviewer disagrees with the automated assessment. Papers rated Average would always have two reviewers. This indication of quality could also be used alongside such measures as citation count to help a reader in prioritising which papers to read first.

Automated recognition of author intentions contained in scientific publication has been successful in the past, as in Argument Zoning (AZ) [17]. Also supporting our idea that author intentions can be linked to better *Related Work* sections is other recent work [7, 14]. These works show that author goals (intentions) identified within a text can be reliably linked to human essay scores. Burstein et al. [3] take this a step further and use discourse analysis to label what they call essay-specific goals, e.g. thesis aim or conclusion. They propose missing labels could be used by students to identify aspects that need improvement in their essay. This relates to our idea that missing author intentions may point to poorer quality material. Whilst these works use the individual labels within their schema to highlight specific missing intentions, our work could be seen as an extension, using the combination of author intentions to suggest an overall indication of quality.

3 Methods

3.1 Author Intentions in Related Work

The author intention labelled data we use is from [5]. They use a data-set from [13] consisting of scientific published papers from the ACL anthology [1]. The labels, based on qualities that Kamler and Thomson [9] have argued should be present in *Related Work*, try to encapsulate neutral citations, those that provide mere description compared to those that highlight gaps or problems, along with identifying where an author talks about their own work and how this relates to the cited work or background in general.

The author intention labels used can be found in Table 1. Certain labels from the original schema were rare and were collapsed into frequent categories. These included sentences positive about a citation/field, works that author's work builds on, uses or is similar to; and comparison of two cited works as described in the description field of Table 1.

Table 1. Related Work Author Intention Labels

Label	**Description**
BG-NE	Description of the state of the field, describing/listing known methods or common knowledge. No evidence i.e. citation is not included
BG-EP	As above but evidence provided i.e.citation included
BG-EVAL	Author highlights a positive or shortcoming/problem or gap in the field
CW-D	Describes cited work or compares two cited works, this could be specific details, or very high level details or nothing more than a reference for further information
A-CW-U	Author's work uses/builds/similar to a cited work
CW-EVAL	A positive or shortcoming/problem or gap about the cited work is highlighted
A-DESC	Author describes their work with no linguistic marking to other's work or being different
A-GAP	Author specifically says they address a gap or highlights the novelty of their work
A-CW-D	Author's highlights how their work is different to cited work
TEXT	Sentence provides information about what will be discussed in the next section

3.2 Assessing Quality

An experiment was set up to rate the quality of each *Related Work* section from the data set in [5]. Participants were presented with the Title, Abstract and *Related Work* section and asked to rate the quality into Poor(P), Good(G) or Average(Avg). Besides this, they were asked (i) if there was enough previous work material; (ii) how well the author related their work to the previous work; and (iii) whether it was clear how the author's work differed from previous work. However, for this work we only use the quality rating given by the participants. Guidance given to participants suggested that it was not enough to list previous work, but that authors should demonstrate the relation of cited work to their own

work. This guidance also indicated that conference papers are usually limited in length so an in-depth explanation of state of the art is not expected.

There were six assessors: four experts and two PhD students – all in the computational linguistics except one student in computer vision. One assessor rated all items, the others rated ten each. Assessor agreement considered the differences between the five assessors and the main assessor who looked at all the articles. Four out of the five assessors were in good agreement with the main assessor; two were in complete agreement and two agreed on seven out of the ten papers. The other assessor only agreed in four instances, which is likely due to them being a PhD student in another area and having less experience with ACL papers. All disagreements were discussed and agreement reached resulting in 50 double rated papers and 44 done by one assessor only. This resulted in a final data set of 94 papers with P-(36%), G-(31%) and Avg-(33%).

4 Mean Label Occurrence in Rated Sections

Table 2. Mean (Var) sentence labels by rating, Significance denoted by * order by Poor/Avg, Avg/Good, Poor/Good

Label	P	Avg	G	Significance
BG-EP	1.2 (0.7)	2 (2)	2.5 (5.1)	* - *
BG-NE	2.2 (10)	3.4 (5.4)	2 (4.5)	- * -
BG-EVAL	0.8 (1.4)	1.4 (3.7)	1.2 (2.5)	- - -
CW-D	8 (46.4)	8 (35.2)	5.6 (20.7)	- - -
CW-EVAL	1.3 (2)	2.3 (5.2)	1.3 (3.2)	* - -
A-CW-U	0.4 (0.3)	0.57 (0.7)	1 (1.3)	- - *
A-DESC	0.5 (0.9)	1.5 (2.4)	1.4 (2.7)	* - *
A-CW-D	0.2 (0.2)	1.2 (1.4)	3.7 (3.7)	* * *
A-GAP	0.1 (0.3)	0.5 (0.5)	1.4 (1.2)	* * *
TEXT	0.2 (0.9)	0.2 (0.2)	0.3 (0.3)	- - -

Table 2 shows the mean number of times a label occurs in each section, grouped by quality rating with variance in brackets. Our intuition is that the occurance of some labels will vary between the different types of ratings. We use Welch's t-test, correct for unequal variances, to test if differences are significant between the means in the groupings. Each group is tested in order of P/Avg, Avg/P and P/G, where * denotes the test was significant ($p < 0.05$).

Our background label with evidence (BG-EP) in our P sections is found to be significantly different to those that occur in Avg or G rated sections. There is a significant difference in the number of background statements in Avg rated sections compared to G sections that provide no evidence (BG-NE). Work is not meant to be cited because it is on the same topic as the citing work, rather it should be cited because it has implications for the author's study [10] and the author should say what these implications are. The findings in Table 2 support this in terms of significant differences between the mean sentences in a G rated

section that describe how the authors work is different to a cited work (A-CW-D) and how the author's work fills a gap (A-GAP). Additionally, we see a significant difference in the number of sentences that describe an author's work (A-DESC) in P rated sections compared to both Avg and G sections.

5 Predicting Quality from Annotated Data

Related Work quality is classified into P, Avg or G. We trained a classifier, experimenting with: SVM (linear kernel), Decision Tree (C4.5) and Linear Logistic Regression(LLR) [6, 12, 16]. We use feature sets of our annotated labels only. Whilst there are many other features that we could include, our focus here is to understand how well our author intentions relate to quality ratings. We use 10-fold cross validation and a majority classifier as our baseline. We report on how our features rank in terms of importance in our best performing classifier.

Table 3. Classifier Performance, Variance in brackets

Classifier	Precision	Recall	Accuracy
LibSVM	0.7 (0.01)	0.7 (0.01)	70 (1.9)
J45	0.6 (0.04)	0.6 (0.05)	57 (5)
Logistic Regression	0.7 (0.02)	0.7 (0.02)	70 (2.2)
Majority Baseline	-	.36	36

Table 4. Ranked Labels-Logistic Regression

	Ranked labels
0.32766	A-CW-Diff
0.21277	A-GAP
0.08085	A-DESC
0.07021	B-NE
0.04681	A-CW-U
0.04468	B-EP
0.00213	CW-EVAL
-0.00851	B-EVAL
-0.01064	TEXT
-0.03723	CW-D

Table 3 shows precision, recall and accuracy from all three classifiers and our majority class baseline. To ensure consistency of results, we ran our models over 10 iterations and report on mean performance (variance in brackets). We test for any differences between our classifiers using corrected t-test, ($p < 0.05$) [11]. All classifiers outperform our baseline significantly. Unsurprisingly, SVM and LLR produce similar results. However, SVM displays marginally less variation in runs, although there is no significant difference between SVM and LLR. Accuracy between SVM and LLR is significantly different to that of the decision tree method. One of the reasons for the latter's poor performance may be that the label features are not exclusive. For example, although author gap and differences (A-GAP, A-CW-DIFF) are rare in P examples, they are not completely absent. We do not have any direct systems to compare to but e-rater 2.0 [4] report agreement between system and human score of essays at 97%. e-rater is, however, a commercial system built on multiple elements not just author intentions. Whilst we do not achieve this level of accuracy, our results are promising

as a first step and with the addition of other features we could improve the accuracy. For example, we experimented with adding sentence counts and citation counts and we were able to consistently improve the accuracy by 4%.

Table 4 ranks labels in terms of importance in SVM, showing that an author highlighting a difference of their work to a cited work or how their work addresses a gap are the most important labels for distinguishing between Quality ratings. This seems plausible as we observe these do occur more in "better" *Related Work* sections. These supports the idea from Maxwell [10] who states that cited work needs to be shown to have implications for the study. It seems that if this type of connection is missing then the work is rated as poorer.

Finally, for our best performing model SVM, we checked the confusion matrix for all 10 iterations. We were interested to see if mis-classification was occurring in the nearest group i.e. G were mis-classified as Avg and not P. We observed that out of 10 iterations this happened twice – one P section being classified as G – and 6 times one G document was classified as P. We speculate that we could improve performance by studying patterns of labels occurring together. When we considered the mean occurrence and variance of labels in Table 2, we saw that it is not simply a case of a P section not having any sentences about the author's work or never mentioning a gap. We believe there may be more to learn about patterns that happen with labels occurring together that support better classification of the different ratings.

6 Conclusions

Using *Related Work* sections, we have shown that some author intentions differ significantly across sections rated P, Avg and G. These author intentions show promise as being viable indicators of quality of the content. We speculate that these different rated sections will have co-occurrence patterns of labels that may provide stronger indications of differences between the quality ratings – an aspect we intend to investigate in the future. Our study does have limitations of the small sample size – 94 papers and only one domain is considered. Our choice of section *Related Work* is also one that does not occur in every domain. Our prediction of quality rating is consistently accurate at 70% with only author intentions as features. Whilst this does not match commercial tool accuracy, such as e-rater (97%), it is a very promising result that could possibly be improved with additional features. Reaching human level of judgement for peer-review in scientific papers is most likely impossible. For example, it is hard to tell what is missing, specifically what has not been addressed or identify something that is incorrect – these aspects might still require a human expert. Nonetheless, we believe that this type of quality rating, if developed at a section specific level, could prove useful in supporting peer-review, directing where reviewers time should be focused and on which papers. In addition, it could help a reader prioritise their reading list of papers.

References

1. Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.Y., Lee, D., Powley, B., Radev, D., Tan, Y.F.: The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In: LREC 2008 (2008), http://www.lrec-conf.org/proceedings/lrec2008/pdf/445_paper.pdf
2. Bridges, D.: Research quality assessment in education: impossible science, possible art? British Educational Research Journal (2009)
3. Burstein, J., Marcu, D., Knight, K.: Finding the WRITE stuff: Automatic identification of discourse structure in student essays. IEEE Intelligent Systems (2003)
4. Burstein, J., Chodorow, M., Leacock, C.: Automated essay evaluation: The criterion online writing service. AI Magazine **25**, 27–36 (09 2004)
5. Casey, A.J., Webber, B., Glowacka, D.: A framework for annotating related works, to support feedback to novice writers. In: Proceedings of the 13th Linguistic Annotation Workshop held in conjunction with ACL 2019 (LAW-XIII 2019). Association for Computational Linguistics, Florence, Italy (Aug 2019)
6. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 (2011)
7. Ghosh, D., Khanam, A., Han, Y., Muresan, S.: Coarse-grained argumentation features for scoring persuasive essays. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 549–554. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-2089
8. Glanzel, W., Debackere, K., Thijs, B., Schubert, A.: A concise review on the role of author self-citations in information science, bibliometrics and science policy. Scientometrics (2006)
9. Kamler, B., Thomson, P.: Helping doctoral students write: Pedagogies for supervision. Routledge (2006). https://doi.org/10.4324/9780203969816
10. Maxwell, J.A.: Literature reviews of, and for, educational research: A commentary on boote and beile's "scholars before researchers". Educational Researcher **35**(9), 28–31 (2006). https://doi.org/10.3102/0013189X035009028
11. Nadeau, C., Bengio, Y.: Inference for the generalization error. In: Proceedings of the 12th International Conference on Neural Information Processing Systems. pp. 307–313. NIPS'99, MIT Press, Cambridge, MA, USA (1999), http://dl.acm.org/citation.cfm?id=3009657.3009701
12. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
13. Schäfer, U., Spurk, C., Steffen, J.: A fully coreference-annotated corpus of scholarly papers from the ACL anthology. In: Proceedings of COLING 2012: Posters. pp. 1059–1070. The COLING 2012 Organizing Committee, Mumbai, India (dec 2012), https://www.aclweb.org/anthology/C12-2103
14. Song, Y., Heilman, M., Beigman, B., Deane, K.P.: Applying argumentation schemes for essay scoring. In: Proceedings of the First Workshop on Argumentation Mining. pp. 69–78. Association for Computational Linguistics (2014), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.672.5185
15. Song, Y., Heilman, M., Beigman Klebanov, B., Deane, P.: Applying argumentation schemes for essay scoring. In: Proceedings of the First Workshop on Argumentation Mining. pp. 69–78. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014). https://doi.org/10.3115/v1/W14-2110, https://www.aclweb.org/anthology/W14-2110

16. Sumner, M., Frank, E., Hall, M.A.: Speeding up logistic model tree induction. PKDD **LNCS 3721**, 675–683 (2005), https://hdl.handle.net/10289/1446
17. Teufel, S.: Argumentative zoning: Information extraction from scientific text. Ph.D. thesis, University of Edinburgh (1999)
18. Walker, R., da Silva, P.R.: Emerging trends in peer review-a survey. Frontiers in Neuroscience (2014)