SIG/MET 2015 Submission – PRESENTATION

Comparative discourse epistemetrics: Research article abstracts and full texts

Bradford Demarest

A rising trend toward linguistic analysis exists in contemporary scientometrics, including the measurement of paper and disciplinary affiliation via semantic profiles of papers (an area of study referred to as semantometrics) (Knoth & Herrmannova, 2014), discerning multiply-authored papers from singly-authored, using stylometric indicators (Rexha et al., 2015), and analyzing the semantic frames of verbs and other terms in citation contexts (Small, 2011; Bertin et al., 2015). Our own current course of research (Demarest & Sugimoto, 2015; Demarest, Larivière, & Sugimoto, 2015) has investigated non-topical terms reflecting the social and epistemic frameworks that differentiate academic disciplines as well as politically partisan cultures (Demarest, 2013), with our previous studies of academic disciplines so far focusing on abstracts for doctoral dissertations as well as for research articles. This quantitative research of social and epistemic context through language use we have previously dubbed "discourse epistemetrics".

Full-text analyses of research articles have recently yielded valuable insights in scientometrics; Bertin et al. (2013) have found marked patterns of citation-oriented verb types used in different sections of research articles, with few differences among disciplines, while Boyack, Small, and Klavans (2013) have reported that including full-text information about the proximity of reference pairs increased the textual coherence of co-citation clusters by up to 30%. Verbs that frame knowledge assertions (including citations as well as assertions of the authors' own knowledge claims) are one of the key components of social and epistemic indicators, based on Hyland's (2005) model of metadiscourse, as they are frequently used to express certainty level and emotional attitude of authors as well as to frame the engagement of the reader. Furthermore, linguistic studies of full-text genres are comparable to other kinds of scientometric studies that derive either implicitly from texts (such as citation-based studies) or explicitly (as do re-citation studies, e.g. Zhao & Strotmann, 2015).

The current work-in-progress seeks to study the full texts of research articles in comparison to abstracts via discourse epistemetrics, with interpretations considering previously conducted analyses of research article and dissertation abstracts. In order to more directly compare the current study's findings to those from Demarest and Sugimoto (2015) and Demarest, Larivière, and Sugimoto (2015), this study will focus on three disciplines – physics, psychology, and philosophy. Research questions follow:

1) Does the support-vector machine-based approach to discourse epistemetrics using common interactive metadiscourse terms as noted by Hyland (2005) yield similar accuracy rates for pairwise comparisons of physics, psychology, and philosophy research article texts as for dissertation and research article abstracts?
2) What features distinguish best between each pair of disciplines for full texts of research articles, and how do these features compare those which best distinguish between sets of abstracts for discipline pairs?
3) What might these differences imply about generic differences between research article texts and abstracts within and across these disciplines?

Methods

Our methodological approach to discourse epistemetrics has been to develop machine-learning models (specifically support-vector machines) based on relative frequencies of terms previously found by Hyland (2005) to be used commonly within academia in ways that are disciplinarily specific. These optimized models are then tested for accuracy on academic disciplines, with the resulting accuracy rate serving as a measure of inverse similarity (i.e., the more accurate the model, the less similar the two disciplines are in their usage of social and epistemic terms). The models are then analyzed on a feature-specific level, considering which features are the strongest indicators of each of the tested pair of disciplines.

For the current study, we seek to sample full-text pre-print articles for three target disciplines from several currently available online repositories: Cogprints (for philosophy, 977 papers are available; for psychology, 1714), as well as arXiv (for physics, which contains 65,183 papers in the unspecified subcategory of physics called simply "physics").

Next Steps

While data for dissertation and research article abstracts are currently in-hand, machine-learning models for full-text research articles for the three disciplines being analyzed are still being processed. Once these are completed, comparisons between full texts and abstracts for the papers sampled here will be undertaken, along with comparisons with previous data for abstracts from research articles and dissertations, considering accuracy as well as specific linguistic features.

References

Bertin, M., Atanassova, I., Larivière, V., & Gingras, Y. (2013). The distribution of references in scientific papers: an analysis of the IMRaD structure. In *Proceedings of ISSI 2013 Vienna*. VIenna, Austria.

Bertin, M., Atanassova, I., Larivière, V., & Gingras, Y. (2015). Mapping the linguistic context of citations. *Bulletin of the American Society for Information Science and Technology*, *41*(2), 26–29. http://doi.org/10.1002/bult.2015.1720410208

Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology*, *64*(9), 1759–1767. http://doi.org/10.1002/asi.22896

Demarest, B. (2013). Measuring identities and differences in epistemic communities in political sub-reddits: A novel machine-learning-based metric. Presented at the METRICS 2013 - ASIS&T WORKSHOP ON INFORMETRIC AND SCIENTOMETRIC RESEARCH, Montreal, Canada.

Demarest, B., Larivière, V., & Sugimoto, C. R. (2015). Coming to terms: A discourse epistemetrics study of article abstracts from the Web of Science. In *Proceedings of ISSI 2015 Istanbul*. Istanbul, Turkey.

Demarest, B., & Sugimoto, C. R. (2015). Argue, observe, assess: Measuring disciplinary identities and differences through socio-epistemic discourse. *Journal of the Association for Information Science and Technology*, *66*(7), 1374–1387. http://doi.org/10.1002/asi.23271

Knoth, P., & Herrmannova, D. (2014). Towards Semantometrics: A New Semantic Similarity Based Measure for Assessing a Research Publication's Contribution. *D-Lib Magazine*, *20*(11), 8–.

Rexha, A., Klampfl, S., Kröll, M., & Kern, R. (2015). Towards authorship attribution for bibliometrics using stylometric features. In *Proceedings of the First Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics*. Istanbul, Turkey.

Small, H. (2011). Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics*, *87*(2), 373–388.

Zhao, D., & Strotmann, A. (2015). Dimensions and uncertainties of author citation rankings: Lessons learned from frequency-weighted in-text citation counting. *Journal of the Association for Information Science and Technology*, n/a–n/a. http://doi.org/10.1002/asi.23418