Using Machine Learning Models To Interpret Disciplinary Styles of Metadiscourse in Dissertation Abstracts

Bradford Demarest
Indiana University
bdemares@indiana.edu

Cassidy Sugimoto
Indiana University
sugimoto@indiana.edu

Abstract

This paper presents the results of a study of disciplinary stylistic differences among dissertation abstracts from physics, psychology, and philosophy. Based on differences in relative frequencies of metadiscourse terms as provided by Hyland (2005), we used a machine learning approach to construct SMO vector support models of each discipline whose average accuracy (88.3%) surpassed a baseline model by 22%. We found that model term weights supported the findings of previous qualitative research regarding differences between disciplines and by extension between hard sciences, social sciences, and humanities. Given the success of the metadiscourse- based model, we conclude by proposing an expanded study to investigate disciplinary style both across disciplines and over time.

Keywords: metadiscourse, disciplinarity, machine learning, support vector model, dissertation abstracts

Introduction

Ken Hyland (2004) notes that academic disciplinary differences are not limited to topicality, but instead reflective of differences in "sanctioned social behaviours, epistemic beliefs, and institutional structures of academic communities" (Hyland, 2004, p.2). Discerning these differences in the writing of disciplines has thus far been mostly limited to qualitative or corpus methods, and has excluded machine learning based methods (a notable exception to this being Argamon, Dodick, and Chase's research (2008), which applied the SMO vector support model to investigate differences in epistemic language in between historical and experimental sciences).

Furthermore, research that has focused on disciplinary style as reflective of disciplinary beliefs and behaviors has entirely excluded dissertations, instead focusing on research articles. To address these two gaps in the research, the current study develops a machine learning based approach to investigate disciplinary style differences, using relative frequencies of metadiscourse terms in the dissertation abstracts of three disciplines: philosophy, psychology, and physics. These metadiscourse terms orient the author to the text itself as well as the reader in establishing epistemological and social norms. The current study's findings support the previous findings of qualitative and corpus-based studies (e.g. Becher, 1987; Hyland, 2008), which established epistemological and social differences among hard sciences, social sciences, and humanities.

Methods

The data used in this study was taken from abstracts for physics, psychology, and philosophy dissertations from the years 1980-1991 contained in the ProQuest dissertation database. Disciplines were operationalized by querying dissertations belonging to at least one subject category containing the string "physics", "psychology", or "philosophy". Abstracts from dissertations with more than one identifying string (e.g. both "physics" and "philosophy") were excluded from the dataset.

Acknowledgements: We would like to thank the National Science Foundation (Grant No. SMA-1208804) for financial support, as well as to ProQuest for sharing dissertation data with us for research purposes. Without their generosity this research could not be conducted.

Demarest, B. & Sugimoto, C. R. (2013). Using machine learning models to interpret disciplinary styles of metadiscourse in dissertation abstracts. *iConference 2013 Proceedings* (pp. 901-904). doi:10.9776/13459

Subsequently, data was divided into modeling data (taken from 1981, 1984, 1987, and 1990), development (from 1982, 1983, 1986, 1989, and 1991), and test data (from 1985 and 1988). Non-consecutive year groupings were chosen to create a model that could capture the evolution of disciplinary style over the decade. The training data set was then balanced for discipline frequencies; the discipline with the lowest number of abstracts was found (philosophy) and abstracts were randomly sampled from the other two disciplines until the training sample contained identical counts of all three disciplines. This yielded 4149 instances, or 1383 instances per discipline. Meanwhile, the test data set was collected as the set of all non-empty abstract records from 1985 and 1988, generating 11625 abstracts (874 of them philosophy, 7550 psychology, and 3201 physics).

For the set of features, a list of 316 words or phrases from six categories expressing interaction from Hyland (2005) was collected, and after removing 13 cross-category duplicates, the resulting 303 terms composed the feature set. These terms express authorial stance toward the text and engagement with the reader. Stance is expressed through hedges (which mitigate certainty), boosters (which amplify certainty), attitude markers (which express authorial affect), and self-mentions (with which the author alludes to herself), while engagement is expressed through imperative verbs and mentions of the reader (via pronouns or phrases like "the reader"). After collecting relative frequencies for the set of 303 terms, the WEKA machine-learning program (Hall et al., 2009) version 3.6.6 was used to create an SMO vector-support model (Platt, 1998) of each discipline in contrast to the other two disciplines (e.g. physics vs. non-physics). Each of these models was then tested against the test data set for classification accuracy.

Results

Table 1 presents the accuracy rate by percentage for each discipline, as well as averaged across all three models. The philosophy model was found to be the most accurate (with a 93.96% accuracy rate), and the psychology model the least (81.92%), but the average (88.3%) still outperformed a baseline classifier using the most-likely category by 22% (the most likely category for each model being non-discipline, generating an average 66% accuracy rate).

Table 1
Accuracy rates (%) for SMO Models

Disciplinary Model	Accuracy (Percentage)
Physics	89.02
Psychology	81.92
Philosophy	93.96
AVERAGE	88.3

Table 2 presents the features from the Hyland's term set that were assigned absolute weights of 2 or more per discipline. In addition to the terms, the table also displays weights (positive valence indicating weighting in favor of a discipline, negative in favor of the non-discipline option in the model) and metadiscourse category to which the term belongs.

Discussion

More interpretation of these results is possible than space allows, but even a brief review reveals telling differences. The positively-weighted features that contribute most strongly to the SMO model of philosophy – "argue", "thought", "claim", "think", "know", "my", "establish", and "true" suggest a field that is like Becher's (1987) description of history: critical, reiterative, and "appealing to the professional judgment of the audience" (Becher, 1987, p. 273). Negatively weighted terms such as "observe", "measure", "increase", "calculate", and "use", along with "we", "known", and "sure", further support this depiction of philosophy, as these counter-terms imply non- philosophy as empirical, quantitative, and communal.

Table 2
SMO feature weights (absolute weight > 2)

Philosophy			Psychology			Physics		
Weight	Term	Category	Weight	Term	Category	Weight	Term	Category
5.1922	argue	H	5.5801	assess	EM	-5.5028	assess	EM
4.5917	thought	B	3.8922	we	SM/EM	5.4482	observe	EM
4.392	claim	H	3.5686	recall	EM	4.7915	calculate	EM
-3.7582	observe	EM	3.3318	would	H	-4.3723	thought	B
3.6938	know	B	3.214	showed	B	-3.9784	argue	H
3.4845	my	SM	-3.2125	calculate	EM	3.9042	agree	AM
-3.4377	measure	EM	-3.0813	argue	H	-3.8457	think	B
3.3882	think	B	-3.0167	observe	EM	-3.7755	refer	EM
-3.0313	increase	EM	-2.7916	agree	AM	3.4943	known	B
-2.9822	we	SM	-2.4078	claim	H	-3.4344	my	SM
-2.6662	calculate	EM	2.3266	indicated	H	-3.386	know	B
2.5346	essential	AM	2.295	suggest	H	-3.3344	would	H
-2.5033	known	B	2.1503	find	B/EM	-3.0695	claim	H
2.3474	establish	B	2.139	appeared	H	-3.058	regard	EM
-2.2566	use	EM	2.1324	likely	H	-2.8161	key	EM
2.1499	true	B	2.1266	typical	H	-2.2775	disagree	AM
-2.0226	sure	B	2.0463	you	EM	-2.2689	recall	EM
			-2.0432	show	EM	-2.2618	?	EM
						-2.2183	indicated	H
						2.2023	determine	EM
						2.1998	allow	EM
						-2.1936	suggest	H
						2.154	estimate	H/EM
						-2.1313	essential	AM
						-2.0246	one's	EM

Note. Category Abbreviations: H = Hedge, B = Booster, EM = Engagement Marker, SM = Self Mention, AM = Attitude Marker.

Psychology is defined by its positive terms including “assess”, “we”, “recall”, “would”, “showed”, “indicated”, “suggest”, “find”, “appeared”, “likely”, “typical”, and “you” as a discipline that is communal, empirical (but, similarly to the way in which Becher (1987) describes sociology, self-conscious about knowledge’s status and methodology), while negative weighted terms (“calculate”, “argue”, “observe”, “agree”, “claim”) indicate the two extremes (one rhetorical and interpretative, the other objectivist and quantitative) between which psychology is positioned.

Physics is most strongly defined, based on positive weights, by “observe”, “calculate”, “determine”, “allow”, and “estimate” which along with other positive weighted terms (“agree”, “known”) suggest a discipline that is, in Becher’s words, “cumulative... tightly structured and atomistic” (Becher, 1987, p. 273), quantifiable and rife with directives for future researchers to expand upon the current work. Negatively weighted terms suggest a non-physics which is more interpretative (“assess”, “refer”, “indicated”, “suggest”) or persuasive (“claim”, “argue”, “think”, “disagree”).

Conclusion

Hyland’s terms serve as a useful feature set with which to model disciplinary voice, achieving reasonable levels of accuracy even when disproportionate distributions of classes exist between training and test data sets. Furthermore, the SMO machine- learning algorithm provides interpretable and insightful information at a term-specific level. That said, the current study has served as a useful pilot in that it has demonstrated a proof of concept. Further optimization of the algorithm and expansion of the feature set of terms to include synonymous terms could lead to even more accurate models, which we

propose to apply in a time-series analysis of major disciplines to analyze style shifts both within and across disciplines over the past century, based on the ProQuest dissertation data.

References

Becher, T. (1987). Disciplinary discourse. *Studies in Higher Education*, 12(3), 261–274.

Argamon, S., Dodick, J., & Chase, P. (2008). Language use reflects scientific methodology: A corpus-based study of peer-reviewed journal articles. *Scientometrics*, 75(2), 203–238. doi:10.1007/s11192-007-1768-y

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations* 11(1), 10-18.

Hyland, K. (2004). *Disciplinary Discourses: Social Interactions in Academic Writing*. Ann Arbor, MI: University of Michigan Press/ELT.

Hyland, K. (2005). *Metadiscourse: Exploring Interaction in Writing*. New York, NY: Continuum.

Hyland, K. (2008). Disciplinary voices: Interactions in research writing. *English Text Construction*, 1(1), 5–22. doi:10.1075/etc.1.1.03hyl

Platt, J. (1998). Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning*. Cambridge, MA: MIT Press.