PRESENTATION

Drawing Into Focus: Methodological Enhancements for Discourse Epistemetrics

A plethora of studies have found that communities that create and evaluate knowledge do so in different ways (Bazerman, 1981; Becher & Trowler, 2001; Cronin, 2005; Hyland, 2000). Knowledge is argued or discovered, observed or assessed, supported through the elegance of a mathematical proof, suggested through the irresistible urgency of an ethical argument, or proven through the reproducibility of cold, hard outcomes. Philosophers make knowledge alone; social scientists may do so in small groups; and physicists do so in vast cohorts.

Informetric studies have looked at community identities and differences based on cited texts, authors, keywords, and topics. However, using social and epistemic language in disciplinary writing as the basis of informetric investigation (what this author has previously christened 'discourse epistemetrics') is an area in need of further development. A previous proof-of-concept study (Demarest & Sugimoto, 2015) set forth a method to discern between pairwise combinations of three widely divergent disciplines – philosophy, psychology, and physics, as applied to dissertation abstracts. Demarest and Sugimoto (2015) employed support vector machines (SVMs) (Cortes & Vapnik, 1995) with linear kernels that used 307 features taken from Hyland (2005) to model disciplinary differences with model accuracies ranging from 86-90%. The current work-in-progress improves and expands upon the method by optimizing machine learning algorithm parameters and expanding the feature set. These improvements improve accuracy rates for inter-disciplinary models of social and epistemic term frequencies by up to 9 percent (95-99%).

Methods

Machine learning algorithms. The current study uses SVMs with linear kernels, per Demarest and Sugimoto (2015), as the SVM algorithm have been shown to perform accurately while providing interpretable feature-level information – i.e., numerical feature weights that indicate both strength and valency of the feature (that is, which discipline is being indicated by a term's increased frequency, and how clearly). In the current investigation, various C parameter values are tested via grid search. All machine learning (including both grid search and SVM modeling for testing) is implemented with the Python Scikit-Learn package (Pedregosa et al., 2011).

Feature expansion. Demarest and Sugimoto (2015) employed a set of 307 non-lemmatized terms taken from Hyland (2005). The current study adapts and expands the feature set by using stemmed features from two studies of Douglas Biber (Biber, 2006; Biber & Finegan, 1989) as well as the previously used Hyland (2005) terms. Biber's terms were chosen due to the similarity between Biber's stance categories and those of Hyland. Hyland (2005) divides interactive metadiscourse into terms that hedge (i.e., mitigate the certainty of an assertion), boost (i.e., amplify certainty), demark affect (i.e., color an assertion to reflect the author's emotional or other evaluative assessment), and explicitly position reader and author in relation to one another. Biber divides stance into affect markers, indicators of certainty and doubt, related hedges and emphatics, and modal terms for necessity, possibility, and prediction. See Table 1 for all feature set term counts.

This core set of terms was then expanded using two electronic lexical databases: WordNet (Fellbaum, 1998) and FrameNet (Baker, Fillmore, & Lowe, 1998). WordNet groups 155,287 nouns, adjectives, adverbs, and verbs into 117,659 synsets (conceptually distinct collections of synonyms). Term sets were derived from WordNet in two ways: (1) direct synonyms for each core term were taken from synsets, and (2) hyponyms, sister terms, antonyms, and troponyms. FrameNet groups over 13,000 terms into roughly 1,200 semantic contexts or frames, based on Fillmore's (1976) Frame Semantics. Term sets were derived from FrameNet by (1) finding each core term's frame and taking all frame elements; and (2) finding frames that fit into the categories of stance (per Biber) or stance and engagement (per Hyland), and taking all frame elements therein.

Sampling. All English-language dissertation abstracts from 1978-2008 from ProQuest with disciplinary tags containing the strings "philosophy", "psychology", or "physics" (but excluding those with more than one of these strings) were stemmed and vectors of relative frequencies for terms were generated, yielding 250,221 texts. Demarest and Sugimoto (2015) accounted for disproportionate samples of disciplines by undersampling, resulting in sample sizes averaging 2075 texts per discipline. The current study instead uses a balanced loss function that accounts for false categorization of minority classes with higher penalties. As such, all texts for each of the disciplines could be used in training or testing models. Finally, all terms in the sample were stemmed in the current study, such that stemmed features could be accurately matched.

Testing method. Optimized models were then generated and tested using nested k-fold cross-validation and a grid search over 20 equally distributed values of the C parameter ranging from 0.001 to 100 for each of the feature sets. Each corpus of two disciplines was divided into ten folds, with nine folds composing the training-development set and the remaining fold serving as the test set. Then, the training and development set is split into five folds; for each value of C, four of these inner folds are used to train an SVM model, which is then tested for accuracy against the fifth. This process is repeated for all inner folds, with accuracy rates averaged. The most accurate C value is then used to train and test a model on the current set of folds of the test set (repeating and averaging over the 10 folds).

Preliminary Results. Table 1 below shows counts of terms per term set as well as preliminary accuracy results for the current study. While all expansions of the core feature set resulted in overall improvement of model accuracy, the WordNet-based feature sets performed the best.

	umber of	ychology-Physics	ilosophy-	ilosophy-Physic
ature Sets	atures	)	ychology (%)	)
re terms	5	93	93	96
re+WN1	41	98	**95**	9
re+WN2	49	**98**	95	**99**
re+FN1	25	98	94	98
re+FN2	9	94	92	96

Table 1: Feature set counts and average accuracy. WN = WordNet, FN = FrameNet. Numerals refer to feature expansion process for the resource as specified above.

Conclusions and Future Work

The current study brings marked improvement to the method of discourse epistemetrics through a combination of machine learning model optimization and feature set expansion. Future studies will investigate the interpretability of terms in the expanded feature sets, and then to seek to test the application of this method to other scholarly texts – specifically article abstracts and full texts, as well as to a wider range of academic disciplines in the hope of exploring underlying structures of disciplinary language reflective of academic kinship as well as divisions.

Bibliography

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1* (pp. 86–90). Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/980845.980860

Bazerman, C. (1981). What Written Knowledge Does: Three Examples of Academic Discourse. *Philosophy of the Social Sciences*, *11*(3), 361–387. https://doi.org/10.1177/004839318101100305

Becher, T., & Trowler, P. R. (2001). Academic Tribes and Territories: intellectual enquiry and the cultures of disciplines (2nd edition).

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers* (Vol. 23). John Benjamins Publishing.

Biber, D., & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text*, *9*(1), 93–124.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cronin, B. (2005). *The Hand of Science: Academic Writing and Its Rewards*. Scarecrow Press.

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. https://doi.org/10.1002/asi.23271

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, *280*(1), 20–32.

Hyland, K. (2000). *Disciplinary discourses: Social interaction in academic writing*. London: Longman.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London: Continuum International Publishing Group.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.