Context matters: on the usage and semantics of hedging terms across sections of scientific papers

Dakota Murray[1], Vincent Larivière[2], Cassidy R. Sugimoto[1]

[1] *dakmurra@iu.edu; sugimoto@indiana.edu*
School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, USA

[2] *vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Canada

Abstract

The electronic availability of the full-text of scholarly publications has made possible large-scale analysis of in-text citations. Many such studies have made use of signal terms—individual words or phrases that suggest some phenomenon of interest. Most studies utilizing signal terms treat them as having stable meaning; however, the usage and meaning of these terms may differ between rhetorical contexts. We conducted a preliminary analysis to investigate the extent to which the usage and semantics of *hedging* signal terms differed between the introduction, methodology, and discussion sections of scientific papers. We sampled three million sentences containing citations from a large database of full-text publications and counted the occurrences of sixteen hedging terms in each section. We found that the incidence of hedges varied across sections with distinct patterns for individual terms. We modelled semantic relationships within a section using word2vec and constructed rudimentary measures to compare the similarity of hedging terms between models trained on each section. We observed that the meaning of hedging terms differed between sections with distinct patterns for individual terms. Our results suggest that signal terms are not independent of their context which has implications for bibliometric full-text analysis.

Introduction

For more than 40 years, sociologists and information scientists have worked towards a better understanding of the function of citations (Moravcsik & Murugesan, 1975; Cronin, 1981; Small, 2004). Early analyses, while insightful, were limited by the lack of accessible data (i.e. electronic journals) and computational tools. However, the increased ubiquity of digital scholarly publications and availably of large-scale databases has made possible sophisticated textual analyses of in-text citations. Some analyses attempted to characterize the contextual attributes of in-text citations, such as their position and distribution within publications (Bertin et al., 2016a; Boyack et al., 2018). Others instead sought to develop machine learning algorithms to automatically classify *citances* (sentences containing citations) according to their function (Teufel, Siddharthan, & Tidhar, 2006), sentiment (Catalini, Lacetera, & Oettl, 2015; Jha et al., 2017), importance (Valenzuela, Ha, & Etzioni, 2015) and more.

A common input feature for citation classifiers is the presence of *signal terms* within a citance. For example, the terms "excellent" and "poor" may prove useful for the task of citation sentiment classification. Other papers have made use of specific dictionaries of signal terms to study more abstract concepts: for example, Small et al., (2017) uses a list of "discovery" terms to track scientific discoveries, whereas Chen et al., (2018) identifies terms that suggest uncertainty. A paper by Di Marco and Mercer (2004) described how a certain type of signal term—*hedges*—might be used to classify citations; however, they also noted how the semantic meaning and usage of hedging terms might differ between sections of a paper. This speaks to a wider issue with the use of signal terms in bibliometric analysis and classification: rhetorical context matters. The meaning or usage of a hedging, discovery, or uncertainty signal term may differ depending on whether the term appears in the introduction, methodology, or discussion. A term is not independent of its context (Bertin et al., 2016b).

In this paper we present a preliminary analysis of the extent to which the incidence and semantic meaning of common signal terms differ between rhetorical contexts of a manuscript. We focus on a particular type of signal term: *hedging* as defined by Hyland (1996), discussed by Di Marco and Mercer (2004), and used in Chen et al. (2018). We consider three distinct rhetorical contexts: the introduction, methodology, and discussion sections. This analysis will inform future textual analyses of scientific publications and establish a foundation for further investigations into how semantics differ between rhetorical and scientific contexts.

Data and Methods

We used data from the Elsevier ScienceDirect database, obtained and managed by the *Centre for Science and Technology Studies* at Leiden University. This data contains full-text for nearly five million English-language full-length articles, short communications, and review articles published between 1980 and 2016. More information about this dataset, including a comparison to the existing PubMed dataset, and a descriptive analysis, can be found in Boyack et al. (2018).

We first sampled 300,000 full-length English-language articles that were published after 2014. From these publications, we sampled approximately one million citances (citation sentences) listed in a section with a name that included the substring "intro"; we repeated this sampling for citances in sections with titles containing the substrings "method" and "discussion" resulting in three corpora. We tokenized sentences within each of these corpora into unigrams and bigrams and tallied the counts and proportions of each token.

We represented the semantic relationships between terms in each corpus using word2vec, a method of training vector space word embeddings from a corpus of text (Mikolov et al., 2013). Each trained word vector represents a single word; words with similar contexts will have high cosine similarities. For example, "researcher" would likely have a similar context as "scientist", and "scholar" and so their vectors would have high cosine similarities. We trained separate word2vec models for each corpus using *genism v3.6.0*. We adhered to the same training parameters as Chen et al., (2018), though we decreased the number of training vectors from 200 to 100 due to our smaller training data; we constructed selective models by including only terms that occur at least 100 times in the corpus.

To compare the meaning of a target term between word2vec models, we need a measure of similarity between models. However, due to the stochastic nature of word2vec training, direct comparison of word vectors between models is essentially meaningless, even when trained on the same data. Some researchers have developed techniques to compare models, but there is little consensus on appropriate technique (Kutuzov, 2018).

We considered two approaches to assess the similarity of a target word's semantic meaning between models. The first approach leverages *global* relationships between the target word and a large sample of words shared between the two models. For two word2vec models we defined their shared common vocabulary. We then measured the cosine similarity between the target term and each word in the shared vocabulary. We used the squared correlation between these two sets of similarities as a measure of semantic similarity. As an additional validity check we trained two word2vec models on an identical corpus of 200,000 abstracts and calculated their global semantic similarity. The smallest value of the 16 hedging terms was 0.983 with a mean of 0.992. That the correlation was high between two word2vec models trained on the same data suggests that this measure of global similarity is relatively robust.

For the second approach we examined only *local* relationships between the target term and a small list of similar words. For a target term, we selected a small number of the most similar words from the introduction word2vec model. We then compared the rankings of these words to the corresponding rankings of the target word in the methods and discussion word2vec models. We created an ad-hoc measure of change in total ranking by dividing the change in rankings between each model by the position of the word in the introduction model similarity list; this index downweighs terms appearing lower in the ranking list. Interpretation of this model's results is less straightforward than for global semantic similarity, but it provides insight into the particular relationships of individual terms.

Results and Discussion

We first assessed the extent to which the incidence of hedging phrases differed across citances by section of scientific articles. Figure 1 shows the counts of the stemmed version of each term within the introduction, methods, and discussion sections. Overall, more hedging terms were used in paper's discussion sections than in either the introduction or methods sections. This difference likely resulted from the rhetorical characteristics inherent to each section, rather than other confounding factors: the total number of citances sampled for the three sections was approximately equal. Moreover, the average number of tokens for a citance in each section differed by at most two tokens (around 28 tokens for methods, and 30 for discussion section).

Whereas hedging tended to be most common in the discussion, perhaps due to the section's typically speculative nature, we also observed differences at the level of individual terms. The most common hedges such as "should", 'may", "indicate", and "might" largely appeared in the discussion section. However, "will" was primarily used in the introduction, whereas "predict" was instead more often used in the methodology. There were also differences in the usage of words between the introduction and methodology: the usage of "should", "must', and "cannot", was roughly similar between the two sections. These findings speak to the heterogeneity of hedging terms—usage of individual terms differed by section.

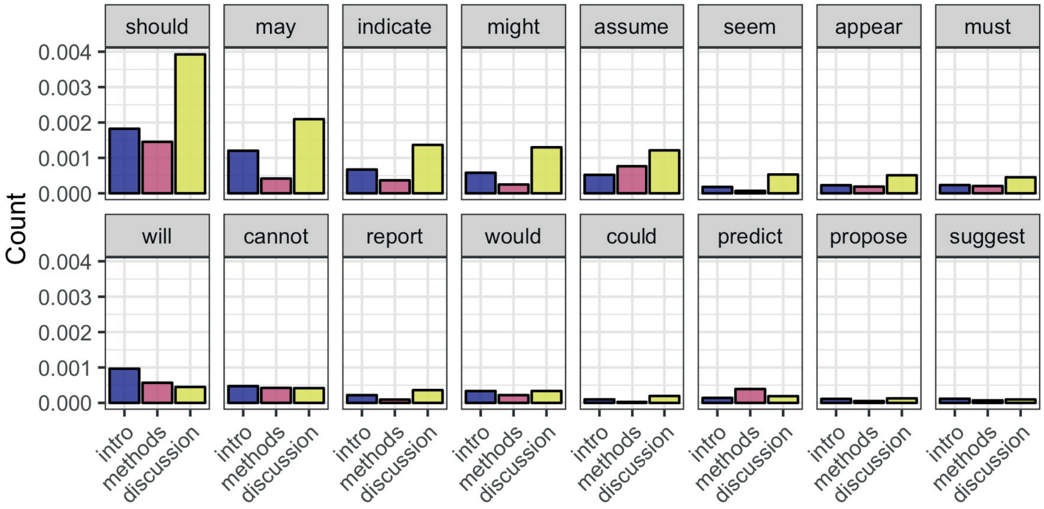

Figure 1. Counts of hedging terms appearing in the introduction, methods, and discussion corpuses. Each corpus includes approximately one million citances. Terms are arranged in order of total instances. These counts include the stemmed version of each term; for example, the count for "indicate" also includes the counts for terms such as "indicates" and "indicated".

In addition to differences in incidence, so too might the *semantic meaning* of individual hedging terms differ by section. For example, "predict" might be used in an introduction to establish a hypothesis: "we predict that *x* is related to *y*". However, in the methods section, this term may instead be used in a more technical context: "we used linear regression to predict *y* given *x*"; in this case, the meaning of the word no longer serves as a hedge in the methodology.

We modelled the semantic meaning of terms by training word2vec models for each section. We investigated the extent to which the semantic meaning of each hedging term differed between each section by calculating an ad-hoc measure of global semantic similarity for each hedging term, and between every combination of two models (Figure 2). We found that, overall, the usage of almost all hedging terms was most similar between the introduction and discussion sections and tended to be least similar between the methodology and discussion sections. Generally, the semantic meaning of hedging terms was similar between the introduction and methods and between the methods and discussion sections.

As in our analysis of the incidence of hedging we also observed heterogeneity in patterns of semantic similarity by individual terms. For example, the term "may" exemplified the general trend of high similarity between the introduction and discussion, but low similarity between these and the methodology section; however, the terms "must", "indicate", and "propose" ran counter to this trend. Model comparisons for the term "must" showed a similar trend as "may", though with a much higher degree of similarity between the methodology and other sections. The semantic similarity of the term "indicate" was instead roughly similar between the introduction and methods, as well as between the introduction and discussion. The term "propose" presented a clear contrast to all other hedging terms, such that the highest semantic similarity was observed between the introduction and methodology sections. This analysis provides evidence that the semantic similarity of hedging differed between sections, and that these patterns of similarity were heterogeneous across individual hedging terms.

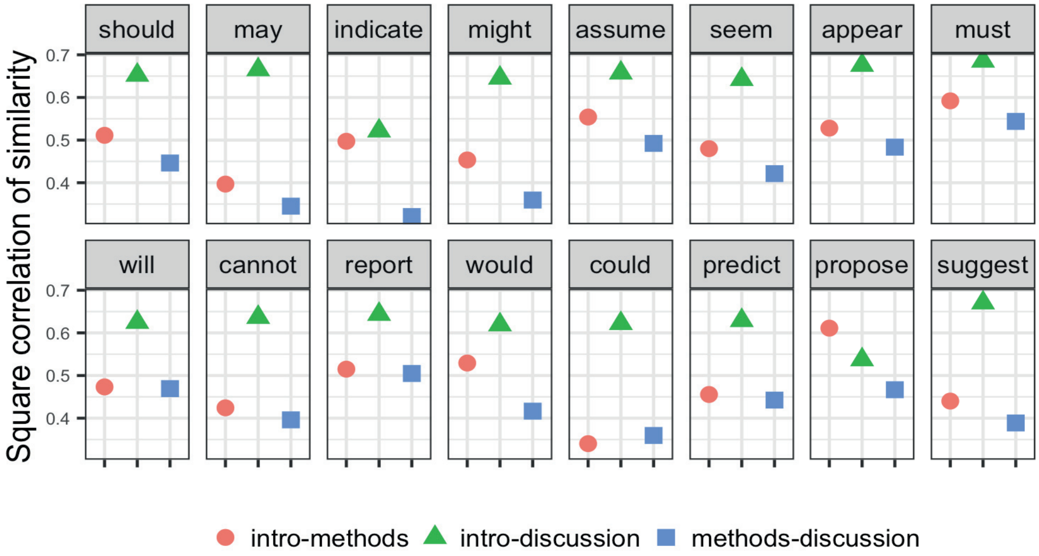

Figure 2. Global term similarity between each hedging term and for each combination of word2vec models. Global term similarity is defined as the adjusted R^2 of the line of best for the top 100,000 most common words shared between both models.

One limitation of this analysis of semantic similarity was that it relies on a measure of *global* semantic similarity, that is, the similarity between a target term and thousands of terms in the shared vocabulary between two models. This measure weighs similarities equally between a target word and all terms in the shared vocabulary. However, a small vocabulary of the most similar words may prove more useful for determining semantic similarity. For example, if we compared the semantic meaning of "dog" across two models, we would be more interested its similarity to "animal" and "pet" than its similarity to "computer". A natural approach is to assess semantic similarity of terms at the *local level*—that is, for a select list of the top *n* most similar terms. Following this approach, we identified the top fifteen most similar terms from the introduction model and traced their change in rankings across the methodology and discussion sections. For this preliminary analysis, we only considered three hedging terms: "indicate", "must", and "propose", each of which exhibited distinct patterns of similarity between the three models. The results of this analysis are shown in table 1.

From this small vocabulary, we observed a mix of both similar and divergent patterns of similarity to those observed in figure 2. Using our global measure of semantic similarity, we found that the term “must” was similar between all three models—here, we also found that “must” had the smallest total change in rankings between the introduction and other models. Moreover, the total change was smaller for the discussion section than for the methods section, potentially corresponding to the higher similarity observed between the introduction and discussion models from figure 2. However, for the term “indicate” we observed roughly similar total changes in rankings between the introduction and methods and discussion models. “Indicate”, with three terms missing in the methods model, also highlights an issue with comparisons between word2vec models—certain terms were excluded because they did not meet the minimum inclusion threshold of 100 instances. Whereas the patterns observed for “indicate” contrasted with our findings from figure 2, those for “propose” were supportive: the total change in rankings terms for the methods model was smaller than the total for the discussion model, though we note that both were especially large compared to other signal terms. One potential explanation for these large changes in rankings may be that “propose” was relatively rare in each corpus. Regardless of comparisons with figure 2, table 1 shows that, even at the local level, the semantics of hedges were heterogeneous between sections.

Table 1. Rank and change-in-rank of three selected hedging terms. For each term, this table lists the top 15 most similar words in the word2vec model trained on introduction citances, measured by cosine similarity. For each term we included the change in rank of each of the 15 similar words between the introduction word2vec model, and the methods (Met) and discussion (Dis) models. Terms not included in a model’s vocabulary are assigned a value of “-”. An ad-hoc change index—the sum of the absolute values of the change in rank, divided by the rank of the term within the introduction model–is shown at the bottom of the table.

Must	ΔMet	ΔDis	Indicate	ΔMet	ΔDis	Propose	ΔMet	ΔDis
should	0	0	reveal	-6	-1	proposes	0	-
will	0	0	demonstrate	-4	-16	we_propose	0	-10
would	-1	0	suggest	-15	2	proposed	-2	-227
might	-4	-4	show	0	-4	put_forward	-9	-172
may	-13	0	reflect	2	0	proposing	-	-
can	-7	2	imply	-	3	presented	-701	-1056
cannot	0	1	highlight	-8	-54	devised	-27	-
do_not	-6	-2	confirm	-141	-11	formulate	-43	-
could	6	2	give	-3	-29	introduced	-37	-829
does_not	-9	-3	relate	-4	-23	gave	-171	-1212
tends_to	-10	-1	give_rise	-	-20	find	-19	-13
they_do	-10	1	find	-33	-40	developed	-185	-676
tend_to	-10	-3	confer	-	-23	we_introduce	6	-
could_potentially	-	5	provide	-22	-7	introduce	-22	-115
able_to	-12	-12	correlate_with	-63	-43	employ	-12	-12
Change Index	10.7	3.8	**Change Index**	41.1	36.1	**Change Index**	169.2	579.5

Conclusion

With this preliminary analysis we found evidence that hedging was not independent of its position within a publication. We observed that hedging terms were more common in the discussion sections of papers, though patterns varied by individual term. We also found evidence that the semantic meaning of hedging terms differed between sections; overall, hedging was most common between the introduction and discussion though distinct patterns were observed for individual terms. We found evidence of this heterogeneity using two distinct approaches. These findings confirm discussions by Di Marco and Mercer (2004) and have direct implications for any study utilizing dictionaries of hedges or other signal terms. These findings also demonstrate that the meaning of a word cannot be fully understood when decontextualized from its context, and that the contextual factors should be considered during textual analysis of scientific publications. The results of this analysis are subject to several limitations. The most important avenue for future research is to further validate our measure of semantic similarity between models and develop more sophisticated techniques. Additionally, this analysis is limited by a relatively small sample of data—one million citances per section—which while large, falls short of the corpus sizes typically used when training more robust word2vec models. Expanding the data size may address some of the issues of terms missing from model's shared vocabulary. Future work will attempt to address these limitations while expanding the scope of analysis to include additional sections (e.g., conclusion, abstract), to distinguish between the semantic meaning of hedging between disciplines, and to examine other common signal terms, such as terms indicating uncertainty (Chen et al., 2018) and discovery (Small et al., 2017).

Acknowledgements

The analyses were carried out by Dakota Murray during a research visit to CWTS. We thank Elsevier for making the data available to CWTS and Nees Jan van Eck for processing. We would also like to thank Ludo Waltman and Wout Lamers from CWTS, and Kevin Boyack from SciTech Strategies, for their feedback.

References

Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V. (2016a). The invariant distribution of references in scientific articles. Journal of the Association for Information Science and Technology, 67(1), 164-177.

Bertin, M., Atanassova, I., Sugimoto, C. R., & Lariviere, V. (2016b). The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. *Scientometrics*, *109*(3), 1417–1434.

Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, *12*(1), 59–73.

Catalini, C., Lacetera, N., & Oettl, A. (2015). The incidence and role of negative citations in science. *PNAS*, *112*(45), 13823–13826.

Chen, C., Song, M., & Heo, G. E. (2018). A Scalable and Adaptive Method for Finding Semantically Equivalent Cue Words of Uncertainty. *Journal of Informetrics*, *12*(1), 158–180.

Cronin, B. (1981). The need for a theory of citing. *Journal of Documentation, 37*(1), 16-24.

Hyland, K. (1996). Writing Without Conviction? Hedging in Science Research Articles. *Applied Linguistics*, *17*(4), 433–454.

Jha, R., Jbara, A.-A., Qazvinian, V., & Radev, D. R. (2017). NLP-driven citation analysis for scientometrics. *Natural Language Engineering*, *23*(1), 93–130.

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1384–1397). Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Marco, C. D., & Mercer, R. E. (2004). Hedging in Scientific Articles as a Means of Classifying Citations.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (pp. 3111–3119). USA: Curran Associates Inc.

Moravcsik, M. J., & Murugesan, P. (1975). Some Results on the Function and Quality of Citations. *Social Studies of Science*, *5*(1), 86–92.

Small, H. (2004). On the shoulders of Robert Merton: Towards a normative theory of citation. *Scientometrics, 60*(1), 71-79.

Small, H., Tseng, H., & Patek, M. (2017). Discovering discoveries: Identifying biomedical discoveries using citation contexts. *Journal of Informetrics*, *11*(1), 46–62.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic Classification of Citation Function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 103–110). Stroudsburg, PA, USA: Association for Computational Linguistics.

Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. Presented at the 29th AAAI Conference on Artificial Intelligence, AAAI 2015, AI Access Foundation.