A Method for Determining the Extent of Recent Temporal Context in Analyses of Complex, Collaborative Thinking

Andrew R. Ruis, University of Wisconsin–Madison, arruis@wisc.edu
Amanda L. Siebert-Evenstone, University of Wisconsin–Madison, alevenstone@wisc.edu
Rebecca Pozen, University of Wisconsin–Madison, rnpozen@gmail.com
Brendan R. Eagan, University of Wisconsin–Madison, beagan@wisc.edu
David Williamson Shaffer, University of Wisconsin–Madison & Aalborg University, david.shaffer@wisc.edu

Abstract: This study presents an empirical method for measuring recent temporal context in collaborative interactions, which can be used to warrant a choice of window length in moving window analyses of complex, collaborative thinking and other interactive learning processes.

Keywords: recent temporal context, moving window, collaborative learning, discourse analysis, epistemic network analysis

Introduction

In the learning sciences, complex thinking is often conceptualized as a process of developing cognitive connections among concepts. In collaborative work, individuals make connections not only within their own contributions but also to the contributions of their collaborators (Shaffer, 2017). To model complex, collaborative thinking, researchers need to identify the context in which such connections are meaningful. Prior work has approached this problem using *moving windows* (Siebert-Evenstone et al., 2016), where each turn of talk in a collaborative conversation is associated with some prior segment of the discussion that forms its *recent temporal context* (Suthers & Desiato, 2012).

A key challenge for window models of complex, collaborative thinking is selecting a window length that is sufficiently long to capture the recent temporal context but not so long as to overrepresent connections that are not meaningful. This study presents a novel empirical method that minimizes the need for human annotation while providing both qualitative and quantitative warrants for determining a window length to analyze collaborative connection-making. We evaluate the method by analyzing conceptual connectivity in the same dataset using different window lengths to explore the effects of window length on the resulting models.

Methods

We analyzed the collaborative interactions of students in the engineering simulation *Nephrotex* (Chesler et al., 2015). In *Nephrotex*, students interact in 4-5 member teams with their engineering advisor through an online instant message program (chat), and the system automatically records all chat conversations for subsequent analysis. *Nephrotex* takes approximately 15 hours to complete. Chat conversations ($N = 54{,}896$ chats) were collected from 20 implementations of *Nephrotex* at five institutions in the United States. Participants ($N = 652$) were first- and second-year college students using *Nephrotex* as part of an engineering course. To measure recent temporal context in this setting, we randomly selected 200 utterances from the 54,896 chats in the *Nephrotex* dataset. For each chat, two independent raters identified all immediately preceding chats in the conversation to which the given chat referred. These annotations indicate the window containing a given utterance and its recent temporal context, where window length is the number of chats from the referring utterance to the earliest referent, inclusive.

To calculate agreement between the two independent raters, we computed Cohen's κ (kappa) for each window length. Kappa was calculated for each window size, x, by assigning a "1" to any utterance that a given rater determined to have window length x and a "0" to all other utterances. Kappa thus indicates the extent to which the two raters agreed in their assessments of which utterances' recent temporal context could be modeled with the same window size. To determine whether kappa scores could be generalized to the population from which they were drawn (> 50,000 chats), we computed Shaffer's ρ (rho) to estimate the expected Type I error rate of kappa given the sample size (Shaffer, 2017). This method (a) allows researchers to empirically determine an appropriate window size by rating a small sample of utterances, and (b) provides a statistical warrant for generalizing from the sample to the population as a whole.

We then tested our technique for empirically determining window length by analyzing a portion of the *Nephrotex* dataset using *epistemic network analysis* (ENA) (Shaffer, 2017). Specifically, we used ENA to model data from two implementations of *Nephrotex* (48 students; 5,757 chats) at window length x for each $x \in \{1, 2, \ldots, 13\}$, and for each window length, we compared the networks of (a) students using an engineering virtual

internship for the first time (novices; $n = 24$), and (b) students using *Nephrotex* after using a different engineering virtual internship (relative experts; $n = 24$).

Results

For all window lengths up to nine, agreement between the two raters was statistically significant for $\kappa > 0.65$ (all $\kappa \geq 0.84$, N = 200, $\rho(0.65) < 0.05$), which indicates that the level of agreement between the two raters would have been $\kappa > 0.65$ for those window lengths had they evaluated the entire dataset. Of the 200 chats examined, 49 (24.5%) made no reference to prior chats, and 51 (25.5%) referenced only the previous chat. However, it is not until a moving window with a length of seven (MW7) that the relevant connections were captured for more than 95% of the sample utterances. No utterance required a window length of more than 18 chats.

To evaluate this method, we computed ENA models at each window length to determine (a) at what window size(s) the model indicates a statistically significant difference between the novices and relative experts that is consistent with our qualitative analysis of the data, and (b) at what window size the ENA metric space stabilizes, such that increases in window size produce no significant changes in the space and thus no changes in model interpretation. While the ability of ENA to discriminate between novices and relative experts is relatively robust to window length, MW7 was one of the best models, and the ENA metric space stabilized at MW7.

Discussion

Our goal was to develop a method that (a) provides both qualitative and quantitative warrants for determining the optimal window length for use in moving window analyses of conceptual connectivity, while (b) minimizing the number of items requiring human evaluation. To assess this approach, two independent raters analyzed a random sample of 200 student chats (< 0.01% of the 54,896 chats in the dataset). This method identified MW7 as the most appropriate window length for analyzing these data. We then constructed ENA models of the data that differed only in the choice of window length. This analysis confirmed that a model with MW7 both (a) provides statistical discrimination between groups hypothesized to exhibit different patterns of conceptual connectivity based on a qualitative analysis and (b) provides a stable interpretation of the ENA model.

As a result, we argue that annotating a subset of data for *furthest referents* makes it possible to analyze recent temporal context and thus determine an appropriate window length to be used in analyses of complex, collaborative thinking. Importantly, this method minimizes the need for human annotation while providing both qualitative and quantitative warrants for choosing a particular window length. While we describe this method by presenting results from one learning context (*Nephrotex*) and one learning analytic technique (ENA), we believe that a similar approach of annotating data by furthest referents will be compatible with different learning settings, different theories of collaborative discourse, and different methods for modeling conceptual connectivity using moving windows. Of course, future research should test our method by repeating this study using other data and models of connectivity. Critically, this method provides a warrant for making generalizations to the population from which the hand-annotated sample was drawn, making it suitable for analyses of complex, collaborative thinking at scale.

References

Chesler, N. C., Ruis, A. R., Collier, W., Swiecki, Z., Arastoopour, G., & Shaffer, D. W. (2015). A novel paradigm for engineering education: Virtual internships with individualized mentoring and assessment of engineering thinking. *Journal of Biomechanical Engineering*, *137*(2), 024701:1-8.

Shaffer, D. W. (2017). *Quantitative ethnography*. Madison, WI: Cathcart Press.

Siebert-Evenstone, A. L., Arastoopour, G., Collier, W., Swiecki, Z., Ruis, A. R., & Shaffer, D. W. (2016). In search of conversational grain size: Modeling semantic structure using moving stanza windows. In C.-K. Looi, J. Polman, U. Cress, & P. Reimann (Eds.), *Transforming learning, empowering learners: The International Conference of the Learning Sciences (ICLS) 2016* (Vol. I, pp. 631–638).

Suthers, D. D., & Desiato, C. (2012). Exposing chat features through analysis of uptake between contributions. In *45th Hawaii International Conference on System Science* (pp. 3368–3377). IEEE.

Acknowledgments

This work was funded in part by the National Science Foundation, the MacArthur Foundation, the Spencer Foundation, the Wisconsin Alumni Research Foundation, and the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison. The opinions, findings, and conclusions do not reflect the views of the funding agencies, cooperating institutions, or other individuals.