WT&D 2025: WINTER TEXT AND DISCOURSE CONFERENCE 2025

PROGRAM FOR MONDAY, FEBRUARY 3RD

Days: next day all days

View: session overview talk overview

09:00-17:00 Optional Activities (Mentorship, Networking, Community-Building)

17:00-18:00 Session 1: Opening Panel on AI and Education Discourse

Owen Henkel: Moderator

Discussants: Art Graesser, Panayiota Kendeou, Sashank Varma

CHAIR: *Owen Henkel*

LOCATION: Alpine-Balsam

Art Graesser, *Panayiota Kendeou* and *Alyssa Wise*
Opening Panel on AI and Education Discourse
PRESENTER: *Art Graesser*

18:00-19:00 Session 2: AI Driven Assessment

LOCATION: Alpine-Balsam

18:00 *Peter Foltz*, *Sam Pugh*, *Chelsea Chandler* and *Brita Elvevåg*
Charactering Dimensions of Thought Disorder with Large Language Models
PRESENTER: *Peter Foltz*

ABSTRACT. Language provides a unique lens to our inner cognitive processes and thereby an indirect window into the brain and potential pathologies. Evaluation in mental health is often performed by clinicians through language-based tasks, such as story retellings, verbal fluency tasks, and structured clinical interviews. Distortions from norms in language serve as indicators of possible disorders. Nevertheless, there are few established standards for measurement and great variability in human ratings. This issue is most particular in the evaluation formal thought disorder, in which disturbances to the structure, organization, or coherence of thought manifest as incoherent speech.

There has been broad research in computational analysis of language in schizophrenia, applying methods such as word count, parsing, word embeddings, and most recently generative AI.

However, generative AI-based methods are stochastic, and may generate different decisions given the same input data. This poses issues for healthcare delivery where lack of consistency in evaluation poses risks for trust by clinicians and variability in diagnosis. In our research, we applied traditional NLP methods and Large Language Models to analyze patient and control transcripts. The methods assessed multiple dimensions of thought disorder and were compared against clinician-based ratings of these dimensions. We further examined the factors that impact consistency and the accuracy of LLM-based predictions. Finally, we examined generated explanations of the model predictions. Results showed that machine-generated ratings matched favorably those of expert humans. Implications are discussed for language features implicated in thought disorder and methods to achieve both high consistency and accuracy using LLM-based approaches.

18:20 *Ziyuan Cao*, *Raj Sanjay Shah* and *Sashank Varma*
Do Large Language Models Perceive the Same Event Boundaries as Humans?
PRESENTER: *Sashank Varma*

ABSTRACT. Humans segment narratives into sequences of meaningful events. Prior research has revealed the indices and metrics that people monitor for evidence of event transitions. Earlier computational modeling using recurrent neural network architectures focused on prediction error spikes at event boundaries. Recent work using Large Language Models (LLMs) such as GPT-3 has found increases in model uncertainty at event boundaries. However, these effects are small in size (i.e., correlations of approximately 0.10), leaving open the question of whether LLMs are sensitive to human-perceived event boundaries in narratives. The current studies extends prior work with GPT-3 to the newer GPT-3.5 and GPT-4 models. We use the “batch” method of prior studies, giving an entire narrative to a model and prompting it to segment the text. This is different than the “incremental” segmentation of narratives by humans, and therefore our second method presents one sentence at a time and after each one prompts the model for whether an event boundary occurred. We correlate the models’ probability judgments of event boundaries with the proportion of humans who perceive event boundaries in the same locations. For batch presentation, GPT-4 correlates more highly with human judgments than GPT-3.5, whereas for incremental presentation, the two models show comparable correlations with human

performance. However, the sizes of the observed correlations are small in size (i.e., approximately 0.14) and vary widely across narratives. We offer reasons for the weak sensitivity of current LLMs to the event boundaries that humans perceive in texts and raise questions for future research.

18:40 *Owen Henkel* and *Bill Roberts*
Using LLMs to help With Story Retell Assessment
PRESENTER: *Owen Henkel*

ABSTRACT. This study explores the potential of Large Language Models (LLMs) in grading reading comprehension assessments, focusing on story retells in Ghana. The research addresses critical challenges in evaluating reading comprehension in Low and Middle-Income Countries (LMICs), bridging text processing, reading comprehension, and educational assessment. Story retell, where students recount a passage in their own words, offers rich insights into comprehension processes. However, grading story retells at scale has been challenging, especially in resource-constrained settings, due to time-consuming manual evaluation and issues with objectivity and standardization. LLMs present an intriguing possibility to address these challenges, potentially combining the benefits of retell assessments with scalability. Our study explores LLM application in grading reading comprehension retells from 13-18-year-old students in Ghana, comparing human rater performance with LLM grading across various rubrics. This approach investigates whether LLMs can offer consistency comparable to human raters while addressing scalability challenges. By exploring LLM performance in LMIC settings, our study contributes to the broader discourse on text processing and comprehension assessment in diverse contexts. It also has the potential to inform educational practices and policies in LMICs, where efficient, scalable assessment tools are particularly needed.

19:00-21:00 Session 3: Poster Session I and Reception

LOCATION: Foyer, Boulderado

Lauren Flynn and *Laura Allen*
A Review of Topic Change During Communication
PRESENTER: *Lauren Flynn*

ABSTRACT. Topics dynamically change over time– influencing the direction of conversation. Studying the complexity of these conversational patterns is crucial to understanding how perceptions of communication

success are formed. However, the study of topic change is challenging due to a wide variety of theoretical lenses, methodological designs, and analytical techniques. For example, the ways in which quantitative and qualitative researchers treat topics differs in how shifts in topics are identified and analyzed. While qualitative researchers use subjective human-coded topic segmentation, quantitative researchers tend to automate this process using computational models. To address these challenges, I plan on discussing findings in the form of a systematic review on how changes in semantic topics are defined and assessed across a variety of sub-fields (e.g., discourse processing, computational linguistics, conversation analysis) and mediums (e.g., speech, computer-mediated conversation, writing). This review will cover how topic transitions are defined (or lack thereof), if they are qualitatively or quantitatively identified (i.e., human-annotated, computer-annotated, or combination of both), how transitions are classified (i.e., are there different types?), and how/if user perceptions are considered (i.e., user-based or researcher-imposed). Understanding these dynamics is essential to better analyze and interpret semantic patterns during communication.

Wesley Morris, *Hanlin Chen*, *Langdon Holmes*, *Joon Suh Choi* and *Scott Crossley*
Using Synthetic Data to Improve the Performance of Automated Summary Scoring
PRESENTER: *Wesley Morris*

ABSTRACT. The use of large language models to automatically score written texts has proliferated in recent years, and automated summary scoring has great potential to provide formative feedback in intelligent texts. However, the accuracy of automatic scoring is limited by the size of the labeled datasets available for training. In this study, we test whether the accuracy of these scoring models can be improved by pretraining them on a large dataset of synthetic summaries (n = 21,906) and source texts (n = 1,826). The topics and sources were generated using Llama 3.0 and the summaries were generated using four different open-source generative language models in order to take advantage of their different datasets and training strategies. We trained two preliminary scoring models, one to predict Content scores and one to predict Grammar scores, on a gold dataset of 4,690 human-scored summaries of 99 sources. We used these preliminary models to assign pseudo-scores to the synthetic dataset. Finally, we tested whether pretraining on the synthetic dataset resulted in increased accuracy compared to models trained only on the gold dataset, using five-fold cross-validation to ensure that differences were not the result of random sampling for the train/valid/test partitions. We found that pre-training on synthetic data improved accuracy for the

model to predict Content scores but had no effect on the model to predict Grammar scores. This finding has implications for the development of automated scoring models that can be used in learning platforms to provide feedback to writers.

Chelsea Chandler, *Rohit Raju* and *Sidney D'Mello*
Enhancing the Cross-Domain Generalizability of Collaborative Discourse Classification with Large Language Models
PRESENTER: *Chelsea Chandler*

ABSTRACT. The automated analysis of classroom collaboration offers valuable insights for educators and learners. However, most tools require models trained on large, labeled datasets tailored to specific contexts, such as curriculum units and user demographics. This is a major hurdle for the practical use of community building support tools as such skills inherently span many contexts. We investigated approaches to enhancing the generalizability of large language models (LLMs) trained to classify three distinct Community Agreements: being committed to the community, moving thinking forward, and being respectful in five diverse datasets spanning various grade levels, demographic groups, and curriculum units, all situated in small group work. We investigated approaches to creating generalizable context-agnostic models of collaboration with either no training data or just a single dataset from one domain. We explored various methods, including traditional fine-tuning, adversarial data augmentation, and zero/few-shot prompting, with RoBERTa and Mistral models. We found that traditional fine-tuning of RoBERTa on a single dataset often led to overfitting to the specific curriculum and language used. In contrast, training with adversarially augmented data significantly improved the model's generalization, with some Mistral implementations outperforming even the best RoBERTa results. Our findings showed that increased model generalizability can be achieved with more thoughtful training techniques and more advanced LLMs. Given the resource-intensive nature of producing labeled datasets for each new classification context, our work offers scalable alternatives for fine-tuning LLMs for collaborative discourse classification in educational settings.

Ekta Sood and *Sidney D'Mello*
Representation Learning for Reading Behaviors: Integrating Cognitive Models with Deep Learning
PRESENTER: *Ekta Sood*

ABSTRACT. Understanding and predicting human reading behaviors is crucial for advanc- ing educational technologies, cognitive science, and AI. Eye movements offer a unique window into processes, such as inference formation and memory consol- idation, during reading. However, existing models struggle to

generalize across reading tasks due to noisy human data and limited availability of large-scale eye-tracking datasets. Additionally, the potential of leveraging advanced deep learning approaches to effectively learn reading behaviors and predict cognitive states, like comprehension, remains largely under-explored. We developed a novel transformer-based network pre-trained on time-series data simulated from the EZ-Reader cognitive model and fine-tuned on real human eye-tracking data. Our representation learning approach introduces the first pre-trained model that can be used to extract rich gaze-embeddings for reading data. We compared our model's eye movement predictions against ablated versions trained solely on hu- man data to evaluate the benefits of cognitive model pre-training. Our approach achieves highest accuracy in predicting eye movement behaviors across three accuracy metrics: Global Fixation Duration, Global Fixation Location, and Global Saccade Amplitude. The best model achieved scores of 0.7177, 0.5832, and 0.8737, showing substantial improvements over baseline models (4%, 12%, and 23%, respectively). These findings demonstrate the potential of bridging cognitive models with deep learning to create robust feature representations for reading data. This novel approach not only advances cognitive science and AI by providing better predictions of eye movements but also lays the groundwork for building adaptive intelligent interfaces to support reading comprehension in real-time, benefiting education and assistive technologies.

Ladislao Salmerón, *Lidia Altamura*, *Laura Gil*, *Amelia Mañá*, *Mario Romero*, *Marian Serrano* and *Cristina Vargas*

Self-regulation of digital reading: effects of a long intervention in high-school students

PRESENTER: *Ladislao Salmerón*

ABSTRACT. The digitization of educational settings has introduced new challenges, particularly in the realm of reading comprehension. Recent meta-analyses confirmed the prevalence of the screen inferiority effect (e.g. Salmerón et al., 2024), i.e. people tend to comprehend slightly poorer digital texts, as compared to equivalent printed texts. A potential explanation relies on the fact that readers tend to adopt a shallow processing style when using digital tools that leads to a miscalibration of their comprehension, which could potentially interfere with self-regulation of their learning (Clinton, 2019). In this study we aimed to improve adolescents' self-regulation of digital reading, by means of a 14-week intervention program. 2005 secondary school students (grades seventh to tenth) from 16 high schools across Spain participated in the program (October 2023 to February 2024) that included weekly reading units. Using a cluster randomized trial at the classroom level, students were assigned to one of four reading conditions: question placement (inserted vs.

post-reading), and feedback type (corrective vs. elaborated). We expected that elaborated feedback and inserted questions, serving as prompts to focus attention on relevant information (McCrudden et al., 2005), would enhance readers' self-reflection. This, in turn, is expected to influence the effort planning for subsequent reading tasks (Zimmerman & Moylan, 2009), fostering higher metacomprehension accuracy. Lastly, we analyzed the potential moderating role of students' prior text comprehension abilities, expecting that students with low comprehension skills may particularly benefit from the intervention. Data are currently being analyzed, and preliminary analyses will be discussed at the conference.

Justin Young and *Charlie Potter*
From Textual Product to Text Production: Engaging Teachers in Writing Process Feedback
PRESENTER: *Justin Young*

ABSTRACT. During the last decade, keystroke logging tools have been developed and utilized to study the writing process, source use, and feedback (e.g., Leijten & Van Waes, 2013; Lindgren & Sullivan, 2019). Further, several recent studies (e.g., (Vandermuelen et al., 2023; Vandermuelen, Van Waes, & Leijten, 2020) suggest that the use of writing process feedback tools can significantly improve educational outcomes for students. However, additional research is needed to understand how to best train and support educators who wish to use writing process feedback software in their classrooms. The research-oriented uses of keystroke logging are well-understood; however, much is still unknown about how usage of writing process feedback tools affects teacher perception and behavior in their classrooms. Building upon previous studies on the potential educational uses of writing process feedback tools, this presentation recounts a case study of the use of the keystroke logging software Inputlog as a tool for teaching preservice English teachers to provide effective writing process feedback on student writing. The presentation will include a demonstration of Inputlog, a description of the instructional intervention, and an analysis of participant feedback.

Dorothea French, *Sidney D'Mello* and *Katharina von der Wense*
Aligning to Adults Is Easy, Aligning to Children Is Hard: A Study of Linguistic Alignment in Dialogue Systems
PRESENTER: *Dorothea French*

ABSTRACT. During conversations, people align to one an other over time, by using similar words, concepts, and syntax. This helps form a shared understanding of the conversational content and is associated with increased engagement and satisfaction. It also affects conversation out comes: e.g., when talking to language

learners, an above normal level of linguistic alignment of parents or language teachers is correlated with faster language acquisition. These benefits make human-like alignment an important property of dialogue systems, which has often been overlooked by the NLP community. In order to fill this gap, we ask: (RQ1) Due to the importance for engagement and satisfac- tion, to what degree do state-of-the-art dialogue systems align to adult users? (RQ2) With a potential application to child language acquisition in mind, do systems, similar to parents, show high levels of alignment during conversations with children? Our experiments show that Chat GPT aligns to adults at roughly human levels, while Llama2 shows elevated alignment. However, when responding to a child, both systems' alignment is below human levels. We further explore lingustic alignment in classroom dialogue, in context with learning outcomes.

Haiyin Yang

Do English Noun Phrase Structures Make Mathematics More Difficult?

ABSTRACT. Reading academic texts has been a perennial challenge for many students. Part of the challenge has to do with the pervasive use of complex noun phrases (CNP) in these texts. In academic texts, writers use CNPs to condense clauses into phrases, create discursive flow, and turn a quality or action into a subject or object participant in sentences (Halliday & Matthiessen 2014). Previous studies (e.g., Fang 2024) have identified reasons for CNP comprehension difficulty, including buried actors and causes, high lexical density, specialized vocabulary, and grammatical metaphors incongruent with everyday human experiences.

This presentation offers another perspective on the comprehension challenges posed by CNPs. Drawing on embodied simulation theory(Bergen 2015), which posits that the mind needs to simulate perceptual, motor, or affective experiences to comprehend the representation of a concept, we show that the positioning of head in a CNP may impact comprehension of the CNP. Specifically, when reading a CNP with multiple layers of abstract heads followed by concrete words, the meaning of abstract words must be retained in the working memory until the processing of concrete words allows the integration of abstract words into the mental simulation. Using English middle school mathematics as a case study, we contrast English CNPs (e.g., the sum of the measures of the interior angles of a polygon) with CNPs in a head-final language, such as Chinese (e.g., a polygon’s interior angle sum). We conclude by arguing that placing more concrete words at phrase-initial positions can facilitate comprehension of mathematics texts.

Christopher Steadman, *Aaron Wong*, *Zhanlan Wei*, *Ryan Baker* and *Caitlin Mills*

Too Hard or Just Right? How Difficulty Affects Student Engagement

PRESENTER: *Christopher Steadman*

ABSTRACT. When presented with challenging material that requires sustained attention, our minds have a tendency to wander quite often. Key theories suggest that these mind wandering episodes may have a negative, cascading effect over time. Based on previous work showing that people go off task more during difficult learning material, we tested the idea that presenting difficult material at the beginning of a lecture will have a stronger cascading effect on attention compared to when difficulty builds linearly over time. Participants were randomly assigned to one of three conditions based on the ordering of the linguistic difficulty of the video materials: Easy, Hard, Medium (EHM; N=55); Hard, Medium, Easy (HME; N=54); and Medium, Easy, Hard (MEH; N=54) and provided responses to mind wandering probes while watching the video. We found that there was a significant interaction between Difficulty Ordering and time (assessed by probe number) in terms of participants' reported mind wandering (TUTs) ($X^2 = 12.58$, $p = .002$). Simple slopes analysis indicated that TUTs significantly increased in both the HME ($p < .001$) and MEH ($p = .04$) conditions as a function of time, but not in the EHM condition. These findings suggest that individuals struggle to recover when their attention drifts, particularly when they are faced with high initial difficulty. This has important implications for content development and highlights the challenges in finding an optimal difficulty level when learning.

Nadina Gomez Merino, *Antonio Ferrer*, *Ana García-Blanco* and *Inmaculada Fajardo*

An eye-tracking study of idiom processing by readers with Autism

PRESENTER: *Inmaculada Fajardo*

ABSTRACT. We investigate how readers with Autism Spectrum Disorder (ASD) process idioms in both oral and written form. Nineteen ASD participants (age range: 12 to 16) without intellectual disability and 20 typically developing (TD) participants matched for age and IQ were asked to read/listen sentences with very familiar idioms (previously piloted) that could either be interpreted as figurative or non-figurative depending on the previous context of the sentence (e.g., break the ice). Idiom comprehension questions' accuracy, question reaction time and eye-movements during sentence reading were registered. Mixed models analyses showed no effects of group, presentation format and idiomaticity on accuracy (above 90% across conditions) but we found an interaction format and idiomaticity for question reaction time: both groups took

longer to answer comprehension questions in the literal than in the figurative condition but only in the written format. Regarding eye-movement (only registered for written format), we found significant interactions between group and idiomaticity: only ASD participants showed significantly longer first pass time, total time and fixation counts in the idiom area when it was preceded by a figurative context than by a literal one. Both groups showed longer idiom regression path for figurative than for literal context. ASD participants seem to read idiomatic expressions more cautiously than literal ones, which paradoxically might have made them respond more quickly to comprehension questions in the figurative condition. Idiom processing theories (direct retrieval vs. compositional accounts, Titone et al., 2019) and previous findings about figurative language processing in ASD will be discussed.
