Automated Evaluation of Written Discourse Coherence Using GPT-4

Ben Naismith, Phoebe Mulcaire, Jill Burstein
Duolingo

BACKGROUND

- Automated writing evaluation (AWE) is commonly used by educational testing organizations
- Discourse coherence is challenging to assess using automated systems
- Prior work with GPT has focused on **holistic** ratings for writing assessment (e.g., Mizumoto & Eguchi, 2023)

DATASET: DET-Coh

- 500 short writing responses from the Duolingo English Test (DET) with human ratings for coherence
- 80 responses double rated to assess human inter-annotator agreement (0.72 exact, 0.93 QWK)

METHODS

- Prompt consisted of a **detailed rubric** for coherence, describing distinguishing features of responses at each CEFR level, followed by few-shot examples (one from each level)
- Compare **rating-only** (single numeric output) vs **rationale-rating** and **rating-rationale** conditions to detect and disambiguate chain-of-thought effects (Wei et al., 2022)
- **Baseline model** using traditional features (e.g. sentence-pairwise overlap) based on Coh-Metrix (Graesser et al., 2004)
- Analyze **generated rationales** for qualitative similarity to human rationales and other patterns

GPT-4 provides ratings for coherence with a **strong level of agreement with human raters**.

Generated rationales identify aspects of responses relevant to human coherence ratings.

The response has variable coherence. A relevant position is presented and a point in favor of using media to learn about different cultures is given. However, the topic is not developed, with the second point about disadvantages ending abruptly so that the intended message is not conveyed. There is some appropriate use of basic discourse markers (e.g., “Firstly”, “Secondly”). The language is sufficient to convey basic ideas on the topic, but lexis from the prompt is used repetitively and at times coherence is affected (e.g., “hardwork that they do in that characters”). Overall, the response produces a satisfactory effect on the reader, negatively impacted by language inaccuracies and lack of development.	The response has variable coherence. The writer begins by stating that it is possible to learn about different cultures through books, movies, and TV shows, and provides some examples of what can be learned. The writer then moves on to discuss the advantages and disadvantages of this approach. However, the response is unfinished and the disadvantages are not fully developed. The response demonstrates some organization, with the use of discourse markers such as “Firstly” and “Secondly” to link ideas. However, there are some inaccuracies in grammar and vocabulary that impact coherence, for example “we acn also get encounter dru”. Overall, the response is somewhat effective in conveying the intended message, but the unfinished nature of the response and inaccuracies in language limit the rating to a 3.

Figure 1. Sample rationales (human, left; generated, right) for the same response

Contact: {ben.naismith,phoebe,jill}@duolingo.com

RESULTS

- GPT-4 ratings strongly outperform a baseline model using traditional coherence features
- GPT-4 ratings exhibit a *central tendency effect* (McNamara et al., 2019) despite not being explicitly trained to minimize error
- Rationales identify coherence-relevant features of responses and look similar to human examples

Metric	Baseline	GPT rationale-rating	GPT rating-rationale	GPT rating-only
Exact	0.36	**0.56**	0.53	0.51
Adj.	0.82	0.96	**0.97**	0.95
QWK	0.39	0.81	**0.82**	0.78

Table 1. Agreement of automatic coherence ratings with human ratings

able to discern some relevant	ideas	but the response is not well
struggles to identify any relevant	ideas	. There is no evidence of
a number of incomplete or incoherent	ideas	, for example , the issue
appropriate for the task and	ideas	are not clearly presented or arranged
The writer expresses two basic	ideas	: that video conferencing applications
appropriate for the task and	ideas	are not clearly presented or arranged
possible to discern some relevant	ideas	, such as the writer’s
appropriate for the task with	ideas	not clearly presented or arranged
the writer expressing two basic	ideas	: that taking notes with pen
possible to discern some relevant	ideas	such as that travel can

Table 2. Concordance of key word “ideas” in generated rationales

DISCUSSION

- Rationales do not provide insight into the *process* of rating, but can highlight useful information for human graders
- Rationales could also inform automatic revision and feedback
- Future work should extend comparison to more recent neural coherence models