Large language model augmented exercise retrieval for personalized language learning

Austin Xu, Klinton Bicknell, Will Monroe

Contact: axu@gatech.edu

Self-directed learning for online language learning

How can we give learners the ability to request content in an online language learning setting?

- Learners should be able to tailor the online learning experience to fit needs

Self-directed learning for online language learning

Can we just use direct similarity search / kNNs?

No!
- Learners describe what they want to learn using “language about language”, e.g., “I want to learn about verbs”
- kNN with embedded learner input returns exercises explicitly about language!
- Semantic gap between learner inputs and exercise content
- Cannot be overcome with large-scale pretraining, e.g., BERT

Problem setup:
- Learners provide text input describing what they want to learn
- Goal: Retrieve the most relevant exercises using a method that is
  - Zero-shot: no relevance labels are available for training
  - Multilingual: exercises are comprised of multilingual sentences

mHyER: Synthesize hypothetical exercises based on user inputs

Experimental results

Retrieval on Tatoeba data
- mHyER outperforms supervised baselines in zero-shot retrieval

Ablations
- Both contrastive learning and generated retrieval candidates contribute to performance gains

References
[2] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In ACL 2023