Title

Towards Idea Discovery in Buddhist Corpora: Extracting and Matching Ideas from Corpora via Idiomatic Expressions

Abstract

Understanding meaning beyond surface form remains a central challenge for large language models (LLMs). In this talk, I present a trajectory of recent and ongoing research on how LLMs engage with nonliteral and context-dependent meaning, focusing on idiomatic expressions as a test case. This work represents a step toward the broader goal of Idea Discovery: enabling systems to extract and cluster ideas or concepts from large text corpora in an unsupervised manner.

These approaches should then be applied in my research for the Intellexus Project, where our goal is to design and integrate NLP-based solutions for the analysis of Buddhist corpora, with a focus on Sanskrit and Tibetan.

I show that LLMs can identify idioms in context, even across languages and with minimal task-specific supervision. However, this apparent success is fragile. Models struggle to reliably link idiomatic expressions to semantically equivalent literal paraphrases. Together, these findings highlight both the promise and the limitations of LLMs as models of conceptual meaning, and they expose a persistent gap between current systems and human-level interpretation.

Bio

Kai is a Computer Science PhD student at Reichman University, Israel, where he also earned his BSc in Computer Science and Entrepreneurship and his MSc in Computer Science. This included, interestingly, 2.5 years of exchange period in TU Darmstadt, Germany, where he also learned German the hard way :)

He focuses on natural language processing (NLP) and is a lead researcher at the Intellexus Project, where he investigates the design and integration of NLP-based solutions for the study of ancient Buddhist corpora in Sanskrit and Tibetan. His work involves training large language models (LLMs), designing and curating downstream tasks and evaluation benchmarks, LLMs' explainability studies, and more.

Kai’s research interests lie in multilingual NLP and low-resource languages, and in addressing linguistic and semantic challenges, such as idiom processing, using LLMs.