Contribution ID: **5** Type: **not specified**

What is Semantic Search Good for in Scholastic Corpora?

Thursday, 5 March 2026 11:30 (30 minutes)

Given the highly intertextual character of scholastic literature, locating the exact source of auctoritates –that is, the authoritative statements commonly evoked in medieval quaestiones –is both a basic, though notoriously demanding, editorial task and an indispensable precondition for more elaborate research. Most available computational solutions aiding this task map direct lexical signals, which, while sufficient to track many cases of reuse, leave out relevant instances of paraphrased or otherwise distorted references. This paper reports on experiments with an alternative approach that relies on similarity search using contextual word embeddings. I will discuss the position of this approach compared to alternative methods (especially the so-called fuzzy search and Retrieval Augmented Generation), highlighting the differences in infrastructural requirements and data model. Focusing on this last aspect, I will discuss the details of implementation that I tested on a corpus of Stephen Langton's Quaestiones Theologiae and a selection of its known sources (Parisian literary production c. 1200). In this, I will argue that semantic search seems to offer a viable solution for middle-sized corpora (~10M words), while being less likely to replace fuzzy search as a primary method of tracking large scale text reuse in sizeable corpora.

Presenter: MALISZEWSKI, Jan (Wydział Filozofii, Uniwersytet Warszawski)

Session Classification: Session 1