

## Natural language processing and information linkage for scholarly digital infrastructures (Keynote)

Speaker: Akiko Aizawa, National Institute of Informatics (NII), Tokyo, Japan

The objective of this talk is to explore how natural language processing can be employed to connect diverse scholarly objects at two different levels of scientific publication: the content level of scholarly documents and the knowledge level, including metadata and the database of scholarly entities. Firstly, at the content level, the use of natural language processing in capturing the semantics of non-linguistic objects in scholarly documents will be illustrated. The most important step at this level is to recognize the document structure in order to associate natural language descriptions with non-linguistic elements such as tables, figures, and mathematical formulae. It is noteworthy that in a scholarly document, layout, logical, and semantic structures are inseparably interlinked. As a result of this entanglement, the structure analysis becomes significantly challenging regardless of whether it is a preprocessing step in conventional approaches or a vision-language understanding task, as is the case in some recent studies. As an illustrative example, this talk will describe our attempt to extract and analyze mathematical formulae in documents. Secondly, at the knowledge level, a brief overview of information linkage will be provided. Starting with the early record linkage problem, integrating distributed bibliographic catalogs has long been a central issue in the operation of digital libraries. Moreover, in recent years, knowledge databases have witnessed rapid growth across scientific disciplines. Natural language processing is used to identify the tuples of scientific entities and their relations based on their semantic interpretation. In particular, coreference resolution and entity linking serve as key techniques for connecting unstructured text to existing knowledge resources. However, it should be noted that despite the long history, several unresolved issues remain. Lastly, the talk will introduce recent advances toward developing a national-level research data infrastructure in Japan.

### Biography

Akiko AIZAWA is a professor at the National Institute of Informatics (NII) and currently serves as the Vice Director-General at NII. Aizawa is also an adjunct professor at the University of Tokyo as well as at the Graduate University of Advanced Studies. Aizawa's research interests include natural language understanding, dialogue systems, text-based content and media processing, and information retrieval. Aizawa has served as an organizer and Program Committee member of related conferences and workshops, and also organized mathematical formula retrieval tasks at NTCIR-10, 11, 12.

*Copyright 2021 for this paper by its authors.*

*Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).*