Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology
Technical University of Munich

NLP-KG: A System for Exploratory Search of Scientific Literature in Natural Language Processing

Tim Schopf and Florian Matthes
{tim.schopf,matthes}@tum.de

ACL 2024 Bangkok, Thailand

Motivation

- Scientific literature on NLP is **growing rapidly.**
- Scientific knowledge on NLP is usually available in large quantities as **unstructured texts**.
- Typically, scientific literature search systems are **limited** to **keyword-based lookup** searches.

Challenging to get an **introduction and overview** of new or unfamiliar scientific fields in NLP.

NLP-KG: A feature-rich system designed to support the exploration of research literature in unfamiliar NLP fields.

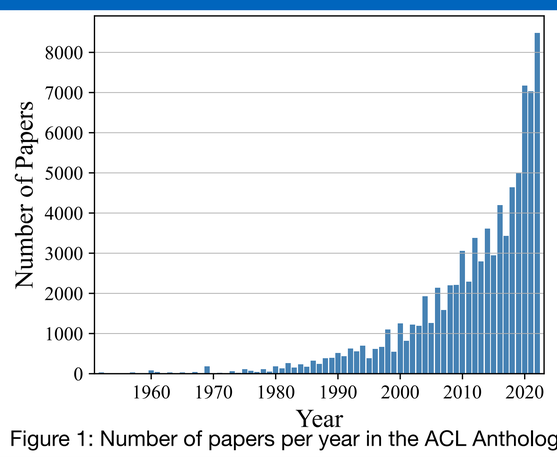

Figure 1: Number of papers per year in the ACL Anthology.

Exploratory Search

- The process of obtaining insights within an **unfamiliar domain** is often directed toward a complex, **open-ended goal**.
- **Exploratory Search** (open-ended goal) ≠ **Information Retrieval** (specific goal)
- Exploratory Search is especially important for researchers:
  - What **approaches** are currently **state-of-the-art**?
  - What are **current** research **trends**?
  - …

Natural Language Processing Knowledge Graph (NLP-KG)

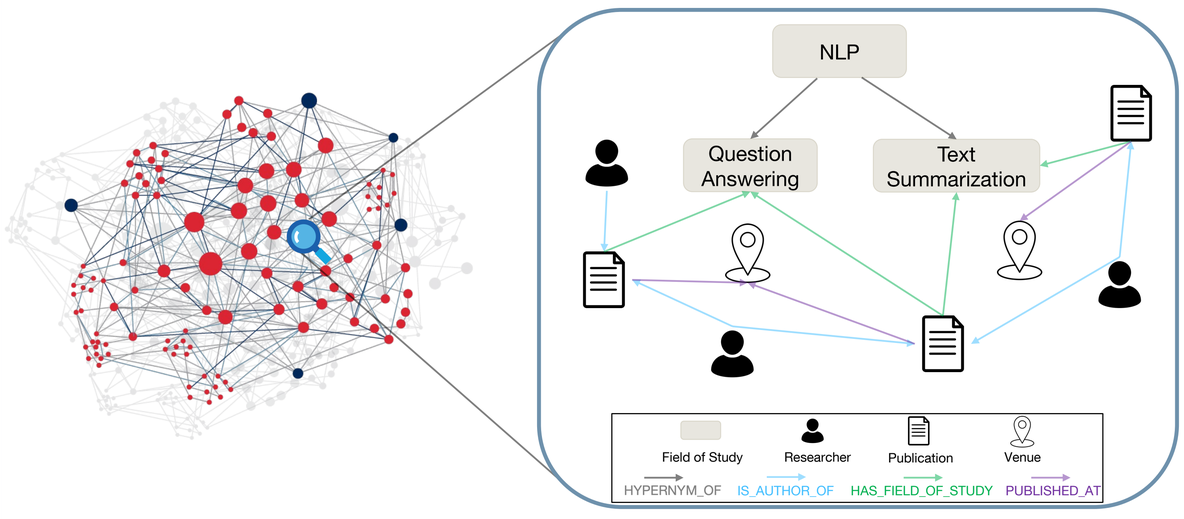

Figure 2: The NLP knowledge graph contains "*Field of Study*", "*Publication*", "*Researcher*", and "*Venue*" entities.

# Fields of Study	# Relations	Max Depth
421	530	7 Levels

Table 1: Overview of the Fields of Study hierarchy developed for the NLP knowledge graph.

System Features

Graph visualization:
- Hierarchically structured Fields of Study in NLP.
- Comprised of (but not limited to) tasks or methods.

Semantic search:
- Keyword-based lookup search for publications, authors, venues, and Fields of Study.

Conversational search:
- Responds to NLP-related user queries in natural language.
- Grounds responses in knowledge from publications using retrieval augmented generation.

Ask this paper:
- Answers user questions about specific publications based on their full texts.

Advanced filters:
- Filter for specific Fields of Study, venues, dates, citation counts, or survey papers.

System Architecture

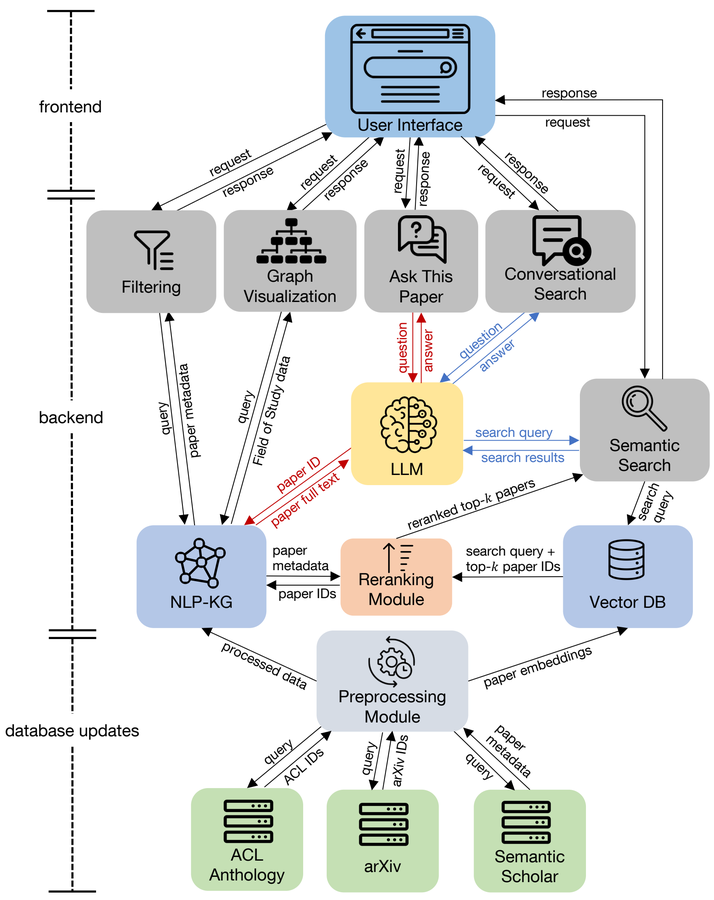

Figure 3: Overview of the system architecture.

- Arrows represent the direction of data flow.
- Red arrows indicate how the autoregressive large language model (LLM) routes the data for the "*Ask This Paper*" feature.
- Blue arrows indicate how the LLM routes the data for the *"Conversational Search*" feature.
- The preprocessing module regularly fetches new publications and processes them to update the knowledge graph and the vector database.

System Evaluation

	Google Scholar	Semantic Scholar	ORKG	NLP Explorer	NLP Scholar	NLP-KG
Keyword-based Search	✓	✓	✓	✓	✓	✓
NLP specific	✗	✗	✗	✓	✓	✓
Fields of Study Tags	✗	✓	✓	✓	✗	✓
Fields of Study Hierarchy	✗	✗	✓	✗	✗	✓
Survey Filter	✓	✗	✗	✗	✗	✓
Ask This Paper	✗	✓	✗	✗	✗	✓
Conversational Search	✗	✗	✗	✗	✗	✓

Table 2: Feature comparison of scholarly literature search systems.

Model	Faithfulness	Answer Relevance
gpt-3.5-turbo-0125	0.9661	0.8479
gpt-4-0125-preview	0.9714	0.8670

Table 3: Evaluation results of the retrieval augmented generation pipeline.

Contact

Tim Schopf
Technical University of Munich
tim.schopf@tum.de

Technical University of Munich
School of Computation, Information and Technology
Department of Computer Science
Chair of Software Engineering for Business Information Systems (sebis)