Knowledge Acquisition Passage Retrieval (KAPR)

Corpus, Ranking Models, and Evaluation Resources

11-09-2024

Artemis Capari, Hosein Azarbonyad, Georgios Tsatsaronis, Zubair Afzal, Judson Dunham, and Jaap Kamps

Knowledge Acquisition Passage Retrieval (KAPR)

ELSEVIER

Task

- Focuses on search in an educational setting, where users seek key educational information

Traditional IR Evaluation

- Focus on topical relevance, often ignoring user-specific relevance
- Binary relevance judgments may include non-informative documents

CLEF 2024 GRENOBLE

Relevance vs Informativeness

Topical Relevance

- *"direct matching between the overall topic of a relevant document and the overall topic of the user need"*

  X Huang, D Soergel (2004)

Informativeness

- Pertinence:
  "Relation between the cognitive state of knowledge of a user and information or information objects (retrieved or in the systems file, or even in existence)"

  (T Saracevic (2006)

Relevance vs Informativeness

Concept in Domain	Snippet	Rel	Info	Explanation
Natural Disaster in Earth and Planetary Science	**Floods are the most frequent natural disaster**. They represent approximately 40% of the total number of natural disasters worldwide. Figure 1 illustrates the number of flood disasters per country from 1990 to 2007. Asian countries carry the largest burden of floods. In particular, India, Bangladesh, and China are the countries where floods affect the most people, on average, during and after the annual monsoon rains. **Floods are defined as natural disasters** when the magnitude of the event exceeds the capacity of a community to anticipate, cope with, resist, and recover from the impact of the natural hazard. Hence, the human consequences of floods depend not only on the severity of the hazard and the level of water, but also on the vulnerability of the affected population…	1	0	Is about ***a sub-type*** of the concept, mainly discussing examples of it without actually explaining it.
Poisson Distribution in Mathematics	**The Poisson distribution arises from** situations where there is a large number of opportunities for the event under scrutiny to occur but a small chance that it will occur on any one trial. The number of cases of bubonic plague would follow Poison: A large number of patients can be found with chills, fever, tender lymph nodes, and restless confusion, but the chance of the syndrome being plague is extremely small for any randomly chosen patient. **This distribution is named for Siméon Denis Poisson**, who published the theory in 1837. The **classic use of Poisson was in predicting the number of deaths** of Prussian army officers from horse kicks from 1875 to 1894; there was a large number of kicks, but the chance of death from a randomly chosen kick was small…	2	1	Is ***about*** the concept, but does not exactly explain the concept.
Narcissism in Psychology	**Hubris as extreme narcissism is egotism, self-centeredness, grandiosity, lack of empathy, exploitation, exaggerated self-love, recklessness, and failure to acknowledge nonmanipulative boundaries** … Hubris lacks basic respect for others. Hubris is often linked with the term “nemesis”, and hints at punishment and suffering resulting from hubristic emotional states of mind (e.g., contempt) toward others. Diminutive states of hubris reflect the classic ,”show-off” personality. Pretentious styles often hide insecurity. They ostentatiously proclaim wished-for minimal or nonexistent assets revealing a sense of deep-seated privation and feelings of inadequacy.	1	2	Not about the concept itself, but rather a sub-concept. Still, very informative for a reader with knowledge of the concept.

KAPR: Test Collection Construction

ELSEVIER

Objective

- Perform Knowledge Acquisition Passage Retrieval (KAPR) task
- Evaluate ranking models
- Correlation relevance and informativeness
- Domain differences

Query Selection

- Select key educational topics/concepts in various science domains
  - 100 concepts across 20 domains
  - 5 concepts per domain: 3 popular, 2 random
- Query format: “What is [concept]?”

CLEF 2024 GRENOBLE

KAPR: Test Collection Construction

ELSEVIER

Document Collection

- **Sources:**

  Review articles from 2,700+ journals and content of 43,000 books

- **Passages:**

  Sections/sub-sections, truncated to max. 500 words

- **Concept-passage corpora:**

  Passages tagged with concepts using XML content processing to reduce search scope

KAPR: Test Collection Construction

Pooling Method

- Dataset: 50 passages per concept
- 5 ranking models, each providing top-50 passages (250 snippets total per concept)
- Final selection based on weighted ranking considering overlap among models

Ranking Models

- Combined results from several models:
  - Lexical search models: BM25, baseline model from user-facing product
  - Semantic search models: 2 bi-encoders, 1 cross-encoder
- Manual comparison of top-3 passages from each model for selection

KAPR: Test Collection Construction

Annotations

- Evaluated by domain experts
  - Each passage reviewed by at least one expert
  - Experts have extensive field knowledge and cross-review for consistency
- Ordinal scale: 0 (not relevant/informative), 1 (partially), 2 (very)

Relevance

- Full coverage and clear discussion of the concept without extraneous information
- Less relevant if it includes other concepts or only specific aspects

Informativeness

- Amount of educational and essential information provided about the concept

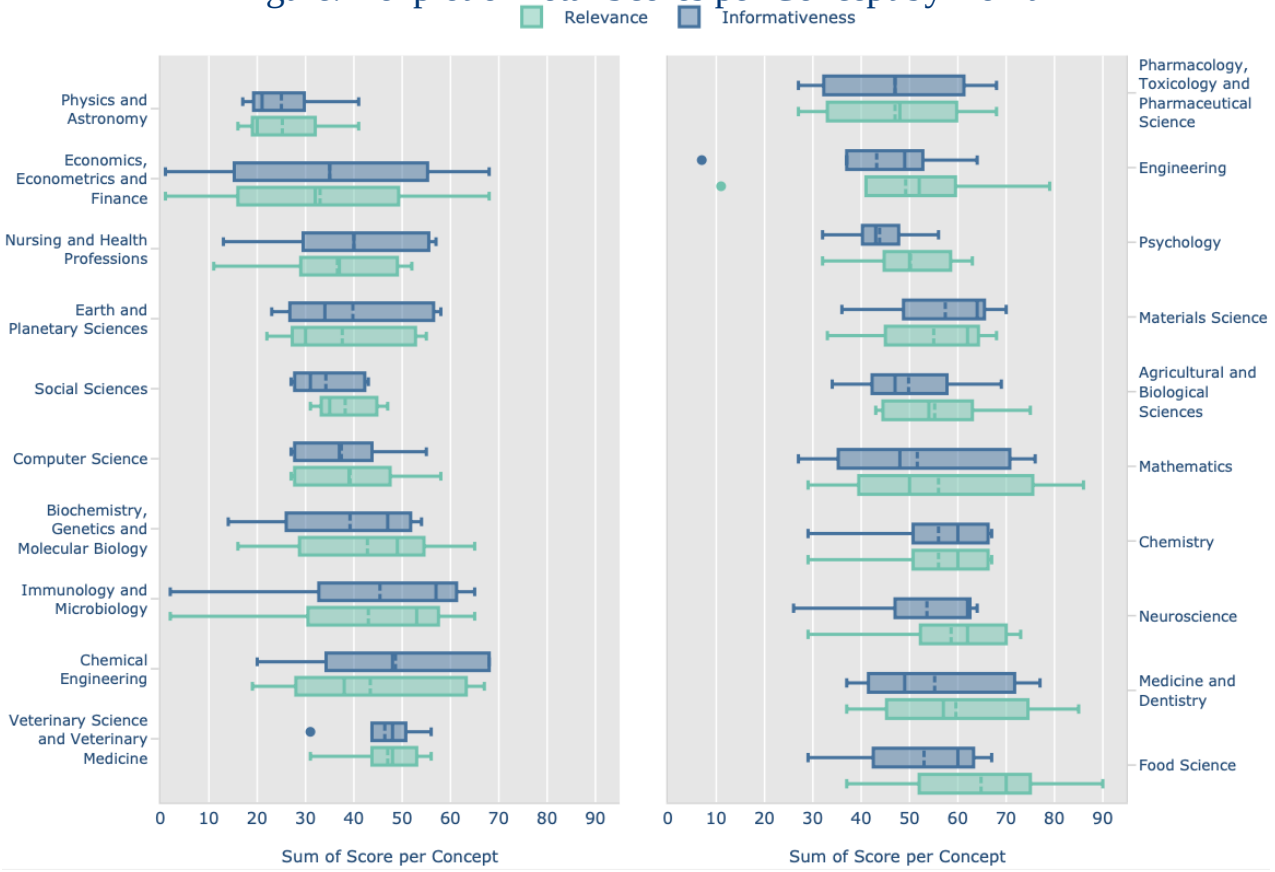
Data Analysis: Relevance vs. Informativeness

- Rel and Info mostly correlated
- ~6% relevant > informative
  - Food Science and Psychology
- ~ 2% informative > relevant
  - Highly detailed but focused on sub-aspects
  - Chemical Engineering, Earth and Planetary Sciences, Nursing and Health Professions

Inf \ Rel	0	1	2
0	31.86%	1.38%	0.04%
1	1.10%	37.84%	4.34%
2	0.02%	1.06%	22.36%

Data Analysis: Domain Variability

Figure: Boxplot of Total Scores per Concept by Domain

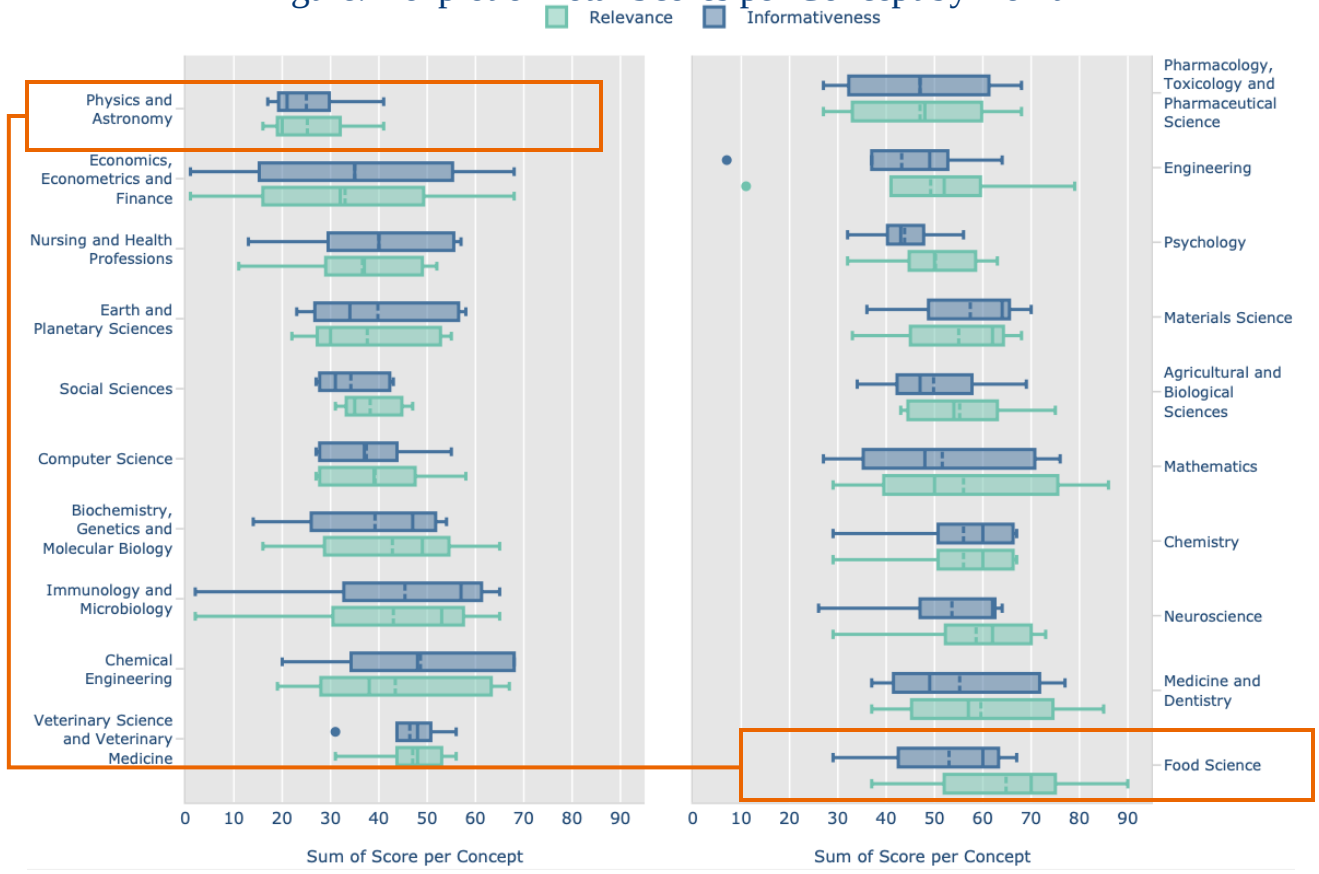


Data Analysis: Domain Variability

Figure: Boxplot of Total Scores per Concept by Domain

Data Analysis: Domain Variability

Figure: Boxplot of Total Scores per Concept by Domain



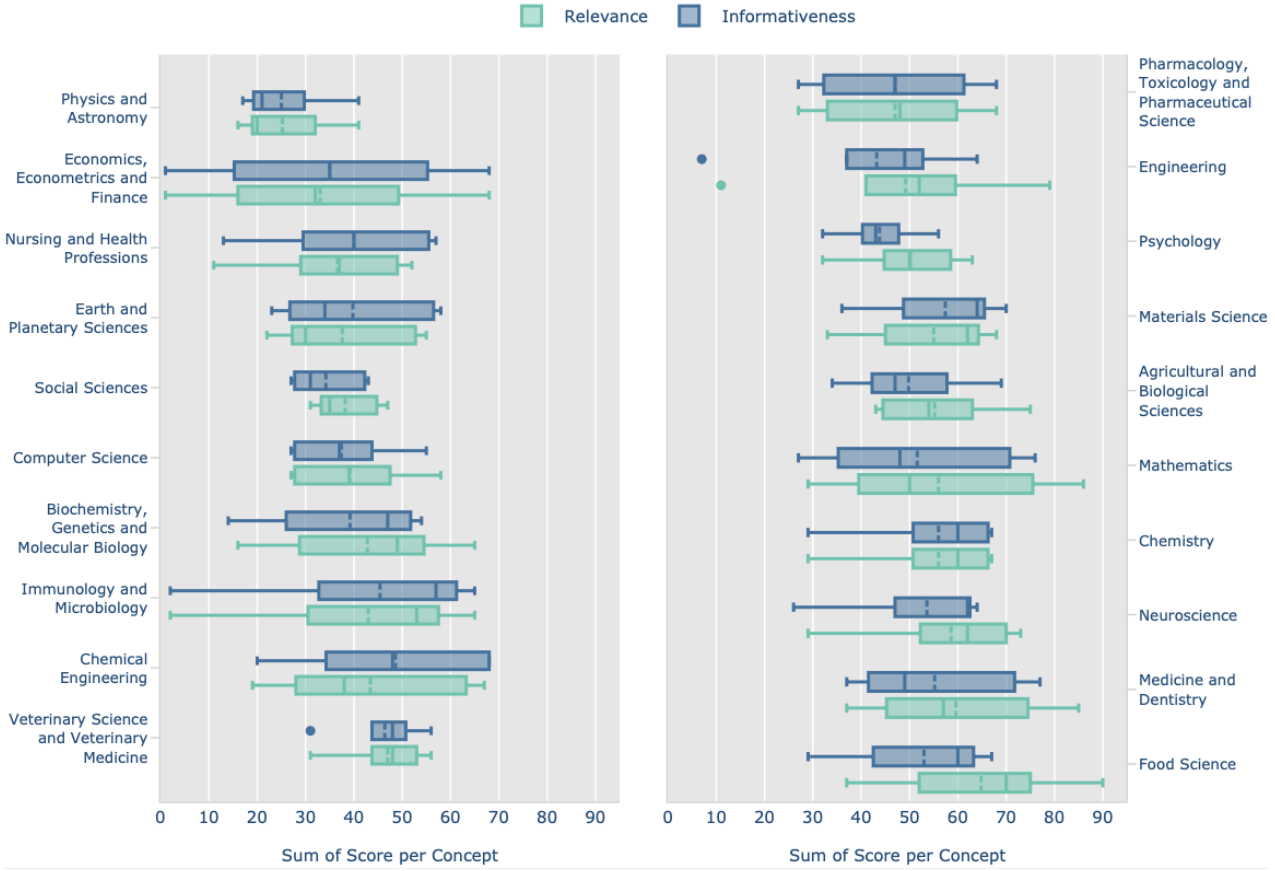
Score Distribution

Scores ↑
Rel-Info Gap ↑

Data Analysis: Domain Variability

Why?
- **Corpus Size**
- **Topic Nature**

Figure: Boxplot of Total Scores per Concept by Domain



Evaluating Ranking Models on KAPR

ELSEVIER

Document Ranking Approach

- **Concept corpus:**

  Initial annotation to find mentions of the scientific concept (query)

- Re-rank documents using ranking models

- Explicit concept mention in context is crucial for comprehension

Evaluating Ranking Models on KAPR

Table: Ranking models used for evaluation

Search Type	Encoder	In Pooling Set	Name
Lexical	-	✓	Baseline Model (TF)
Lexical	-	✓	BM25
Semantic	Bi-Encoder	✓	ST msmarco-distilbert-base-tas-b
Semantic	Bi-Encoder	✓	ST msmarco-distilbert-base-v4
Semantic	Bi-Encoder	✗	ST msmarco-bert-dot-v5
Semantic	Bi-Encoder	✗	ST msmarco-MiniLM-L-6-v3
Semantic	Bi-Encoder	✗	ST RoBerta-large-v1
Semantic	Bi-Encoder	✗	flax-distilRoBerta-v3
Semantic	Bi-Encoder	✗	ST T5-xl
Semantic	Bi-Encoder	✗	ST gtr-t5-l
Semantic	Cross-Encoder	✓	ST msmarco-MiniLM-L6-v2
Semantic	Cross-Encoder	✗	ST msmarco-Electra
Semantic	Cross-Encoder	✗	ST ms-marco-MiniLM-L-12-v2

Model	Aspect	P@10	R@10	R@50	MRR@10	nDCG@10	nDCG@50
BASELINE	rel	0.52±0.09	0.17±0.06	0.36±0.07	0.84±0.08	0.46±0.09	0.40±0.06
	info	0.51±0.09	0.17±0.06	0.36±0.07	0.85±0.08	0.45±0.08	0.40±0.06
BM25	rel	0.54±0.13	0.18±0.06	0.51±0.11	0.82±0.10	0.46±0.10	0.49±0.10
	info	0.54±0.13	0.179±0.06	0.51±0.11	0.82±0.10	0.46±0.09	0.48±0.10
BE ST msmarco-distilbert-tas-b	rel	0.77±0.11	0.25±0.06	0.78±0.06	0.91±0.11	0.69±0.10	0.75±0.06
	info	0.76±0.10	0.25±0.06	0.78±0.06	0.92±0.11	0.66±0.07	0.74±0.06
BE ST msmarco-distilbert-v4	rel	**0.78±0.11**	**0.26±0.06**	0.80±0.05	0.92±0.10	**0.70±0.09**	**0.77±0.05**
	info	**0.78±0.10**	**0.26±0.06**	0.81±0.05	0.91±0.09	**0.68±0.07**	**0.76±0.05**
CE ST msmarco-MiniLM-L6-v2	rel	0.76±0.11	0.25±0.06	**0.82±0.05**	0.93±0.11	0.68±0.09	0.77±0.05
	info	0.75±0.10	0.25±0.05	**0.81±0.04**	0.93±0.12	0.65±0.08	0.75±0.05
CE ST msmarco-MiniLM-L12-v2	rel	0.76±0.10	0.25±0.06	0.75±0.05	**0.94±0.09**	0.69±0.08	0.73±0.05
	info	0.76±0.10	0.25±0.05	0.74±0.05	**0.93±0.09**	0.67±0.07	0.72±0.04
CE ST msmarco Electra	rel	0.70±0.12	0.23±0.06	0.66±0.06	0.90±0.11	0.64±0.09	0.66±0.05
	info	0.79±0.12	0.23±0.05	0.66±0.06	0.90±0.10	0.62±0.09	0.65±0.05
BE ST T5-xl	rel	0.66±0.09	0.21±0.05	0.62±0.09	0.87±0.11	0.60±0.07	0.62±0.06
	info	0.66±0.08	0.21±0.05	0.62±0.09	0.87±0.11	0.59±0.06	0.61±0.06
BE ST gtr-t5-l	rel	0.74±0.09	0.24±0.06	0.69±0.08	0.92±0.11	0.67±0.09	0.69±0.07
	info	0.74±0.10	0.24±0.06	0.69±0.09	0.92±0.09	0.65±0.08	0.68±0.07
BE ST msmarco-bert-dot-v5	rel	0.63±0.13	0.21±0.04	0.65±0.07	0.82±0.16	0.54±0.11	0.61±0.06
	info	0.62±0.11	0.20±0.04	0.65±0.07	0.82±0.17	0.50±0.08	0.59±0.05
BE ST RoBerta-large-v1	rel	0.64±0.09	0.22±0.06	0.60±0.10	0.86±0.10	0.58±0.09	0.60±0.08
	info	0.64±0.09	0.22±0.06	0.60±0.10	0.87±0.10	0.57±0.08	0.60±0.08
BE flax-distilRoBerta-v3	rel	0.66±0.11	0.22±0.06	0.62±0.09	0.86±0.12	0.58±0.09	0.61±0.07
	info	0.66±0.10	0.22±0.06	0.63±0.09	0.85±0.11	0.57±0.07	0.60±0.07
BE ST msmarco-MiniLM-L-6-v3	rel	0.71±0.11	0.23±0.06	0.68±0.08	0.90±0.13	0.63±0.10	0.66±0.07
	info	0.69±0.11	0.23±0.06	0.68±0.08	0.89±0.13	0.59±0.07	0.65±0.05

Model Performance

- Semantic models > lexical search models
- Top model: BE ST msmarco-distilbert-v4
- Small performance gaps among semantic models; no significant difference between cross-encoders and bi-encoders
- Precision and Recall show minimal differences due to binary labeling
- Potential bias in evaluation due to pooling models (R@50 scores indicate bias)
- nDCG@10: All models perform better on relevance than informativeness
- High standard deviations in domain performance → poor generalization

Domain-Specific Performance

Table: Best and worst performing domains for various models

	CE ST MiniLM-L6-v2	BE ST TAS-B	BE ST distilbert-v4	BM25
Top-3 Best Domains	Neuroscience	Agricultural and Biological Sciences	Biochemistry, Genetics and Molecular Biology	Economics, Econometrics and Finance
	Chemistry	Neuroscience	Food Science	Social Sciences
	Medicine and Dentistry	Mathematics	Neuroscience	Immunology and Microbiology
Top-3 Worst Domains	Immunology and Microbiology	Immunology and Microbiology	Chemical Engineering	Chemical Engineering
	Nursing and Health Professions	Chemical Engineering	Immunology and Microbiology	Engineering
	Earth and Planetary Sciences	Earth and Planetary Sciences	Nursing and Health Professions	Earth and Planetary Sciences

Conclusion

Is Topical Relevance enough?

- Relevance judgments often criticized for superficiality and mere topicality
- Existing IR benchmarks may not suffice for KAPR tasks
- Datasets cover limited scientific domains

KAPR Dataset

- Addresses relevance and informativeness criticisms
- Covers wide range of domains
- High relevance-informativeness correlation, varying by domain

Conclusion

Model Performance

- Semantic models > lexical models
- Lexical models more consistent in relevance and informativeness
- Semantic models excel in relevance due to traditional IR training

Challenges and Future Work

- Performance differences across domains
- Single annotator per document introduces bias
- Semantic models struggle with certain domains → poor generalization

Knowledge Acquisition Passage Retrieval (KAPR)

Corpus, Ranking Models, and Evaluation Resources

11-09-2024

Artemis Capari, Hosein Azarbonyad, Georgios Tsatsaronis, Zubair Afzal, Judson Dunham, and Jaap Kamps