Sabine Wehnert
Visakh Padmanabhan
Ernesto William De Luca

Hybrid Legal Norm Retrieval:

Leveraging Knowledge Graphs and Textual Representations

LEIBNIZ INSTITUTE FOR EDUCATIONAL MEDIA
|Georg Eckert Institute
OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG
INF FAKULTÄT FÜR INFORMATIK

Agenda

01 Motivation

Explainability in Information Retrieval

02 Related Work in the COLIEE Competition

Explainability of submissions for Task 3 in COLIEE for the past 5 years

03 Retrieval System

Hybrid relevance scoring

04 Results and Conclusion

Performance vs. Explainability

01

Motivation

Explainability in Information Retrieval

Motivation

- **Prevalence of LLM-based retrieval models** in recent editions of the Competition on Legal Information Retrieval and Entailment (COLIEE)
- Problem: **low transparency** due to black-box models

Should a high-performing yet unpredictable method be preferred over a more explainable but probably less accurate one?

02

Related Work

Explainability of submissions for Task 3 in COLIEE for the past 5 years

COLIEE Task 3 (Statutory Law Retrieval)

Article (Premise)

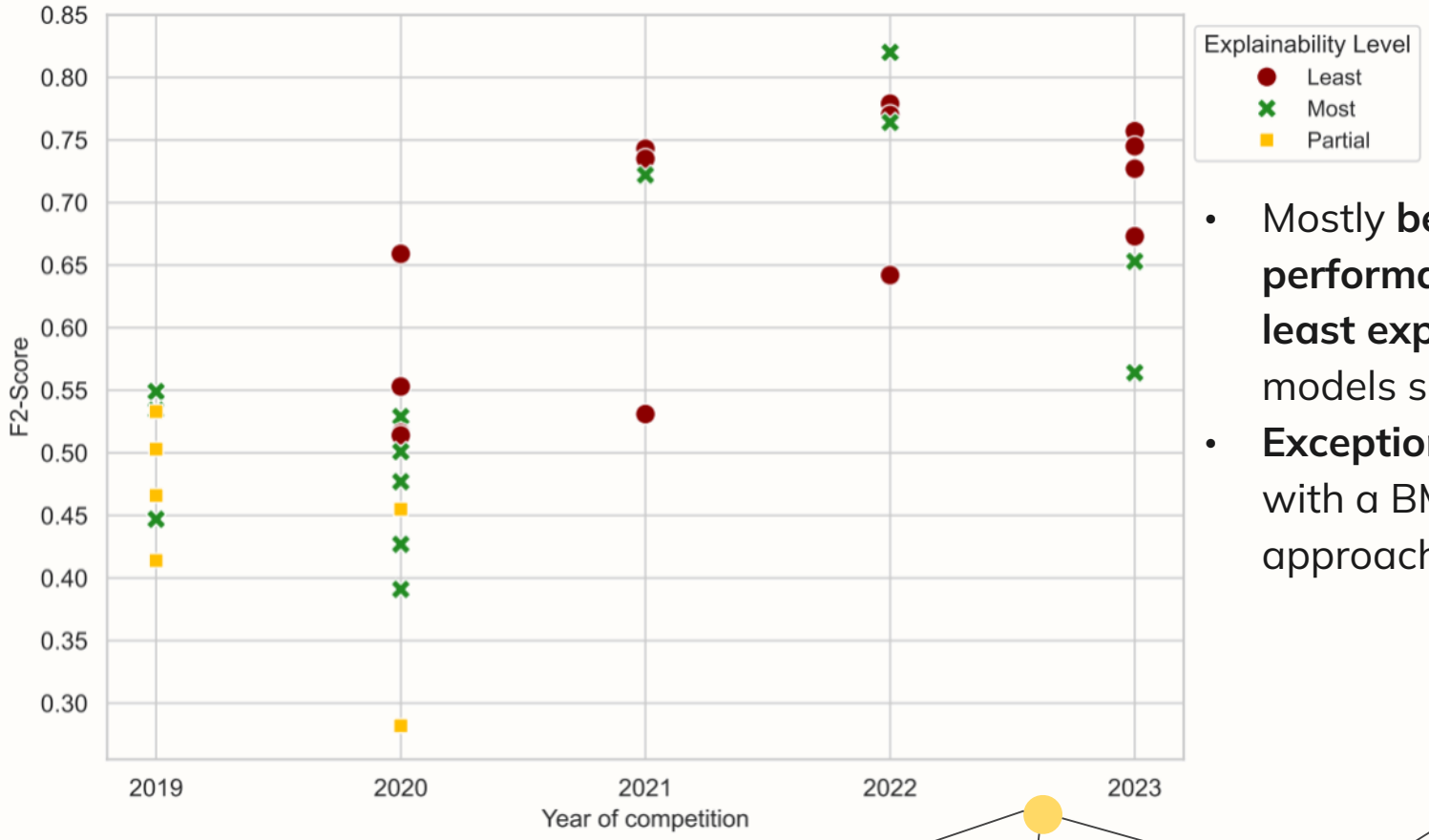
(Seller's Warranty in cases of Superficies or Other Rights) Article 566

(1) In cases where the subject matter of the sale is encumbered with for the purpose of a superficies, an emphyteusis, an easement, a right of retention or a pledge, if the buyer does not know the same and cannot achieve the purpose of the contract on account thereof, the buyer may cancel the contract. In such cases, if the contract cannot be cancelled, the buyer may only demand compensation for damages.
[...]

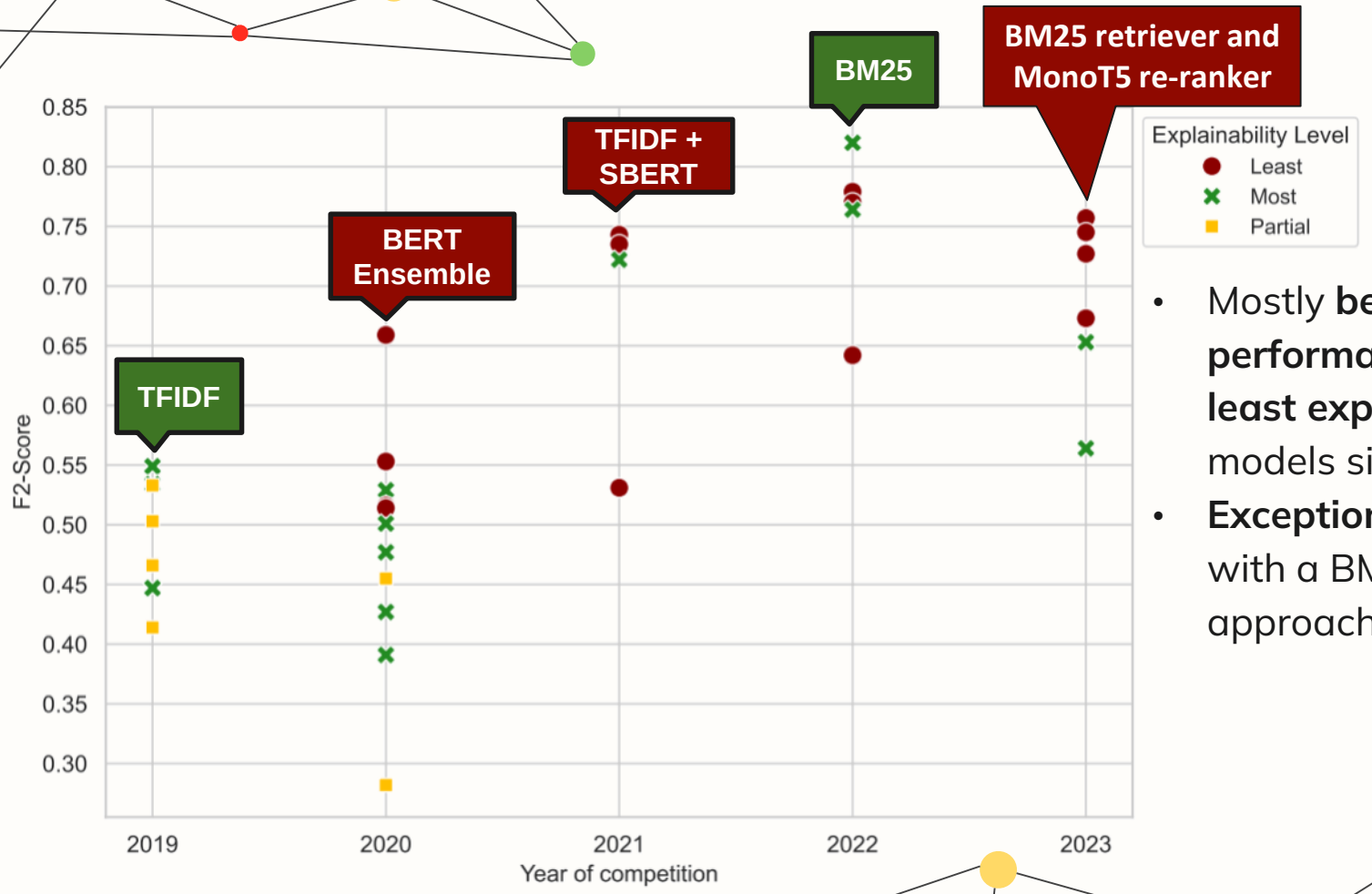
Query (Hypothesis)

There is a limitation period on pursuance of warranty if there is restriction due to superficies on the subject matter, but there is no restriction on pursuance of warranty if the seller's rights were revoked due to execution of the mortgage.

https://sites.ualberta.ca/~rabelo/COLIEE2024/



- Mostly **best performance by least explainable** models since 2020
- **Exception:** HUKB with a BM25-based approach in 2022



2024: BERT-base Japanese + LLM re-rankers

- Mostly **best performance by least explainable** models since 2020
- **Exception:** HUKB with a BM25-based approach in 2022

03

Retrieval System

Hybrid relevance scoring

Relevance Scoring

hybrid score(article,query) =
article similarity

Relevance Scoring

hybrid score(article,query) =

w_{art} · article similarity

+ w_{comm} · commentary similarity

+ w_{prec} · precedent similarity

+ w_{cont} · context similarity

+ w_{graph} · hops score

Relevance Scoring

hybrid score(article,query) =

w_{art} · article similarity

+ w_{comm} · commentary similarity

+ w_{prec} · precedent similarity

+ w_{cont} · context similarity

+ w_{graph} · hops score

- Between 0,1 (weights add up to 1)

Relevance Scoring

hybrid score(article,query) =

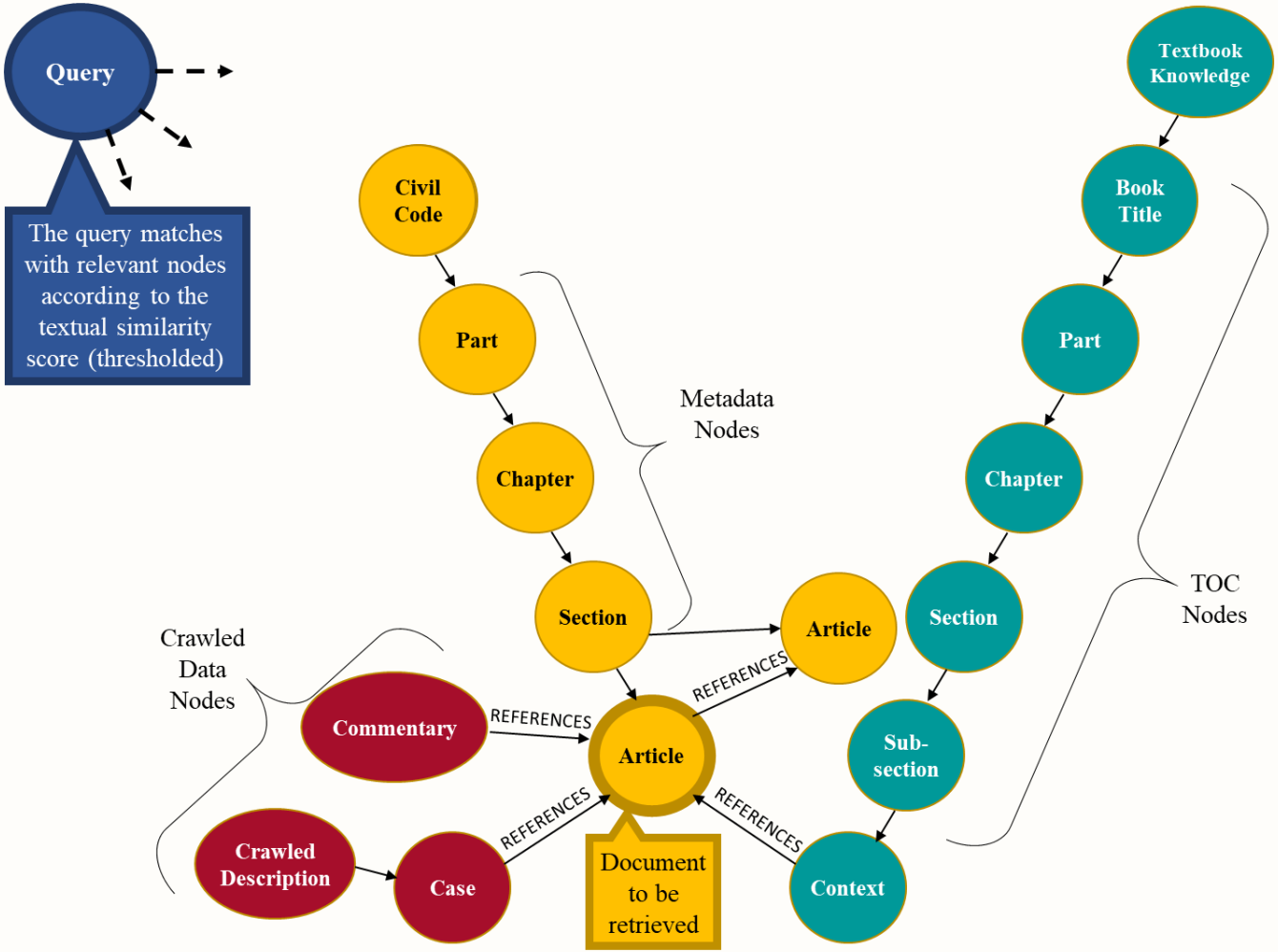
w_{art} · **article similarity**
+ w_{comm} · **commentary similarity**
+ w_{prec} · **precedent similarity**
+ w_{cont} · **context similarity**
+ w_{graph} · **hops score**

Textual Similarity

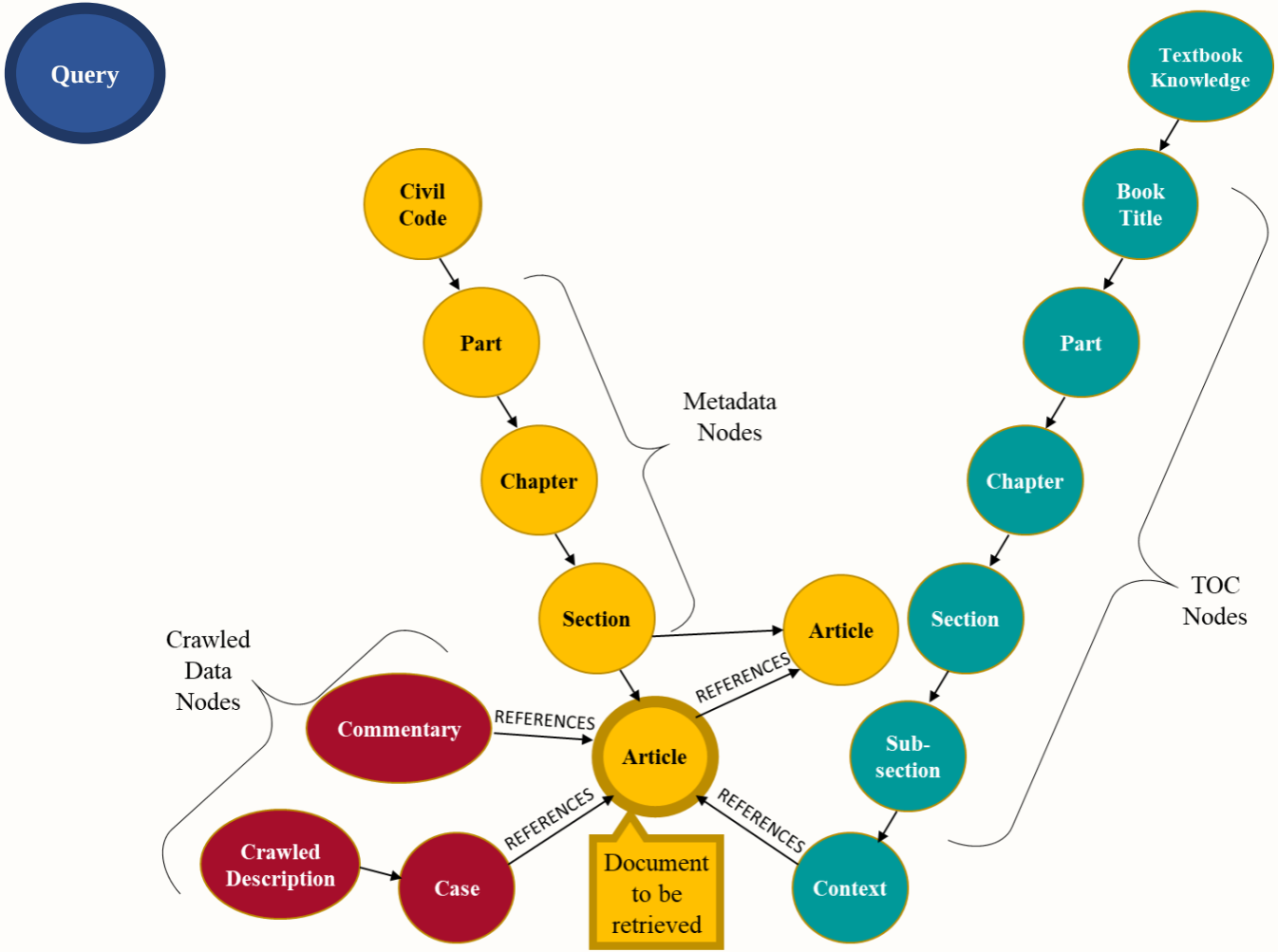
Graph-based Distance

- Between 0,1 (weights add up to 1)

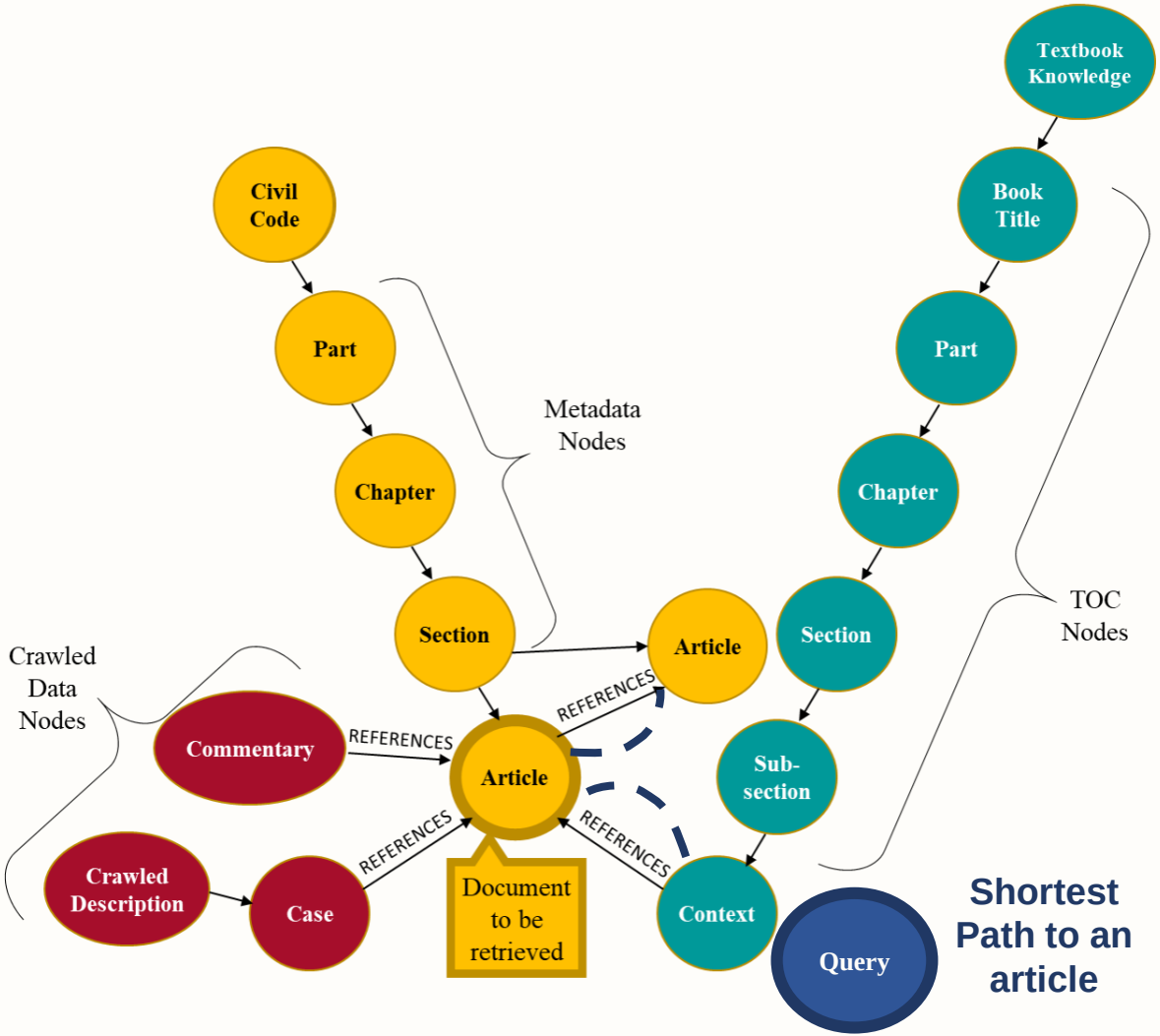
Graph-based Distance



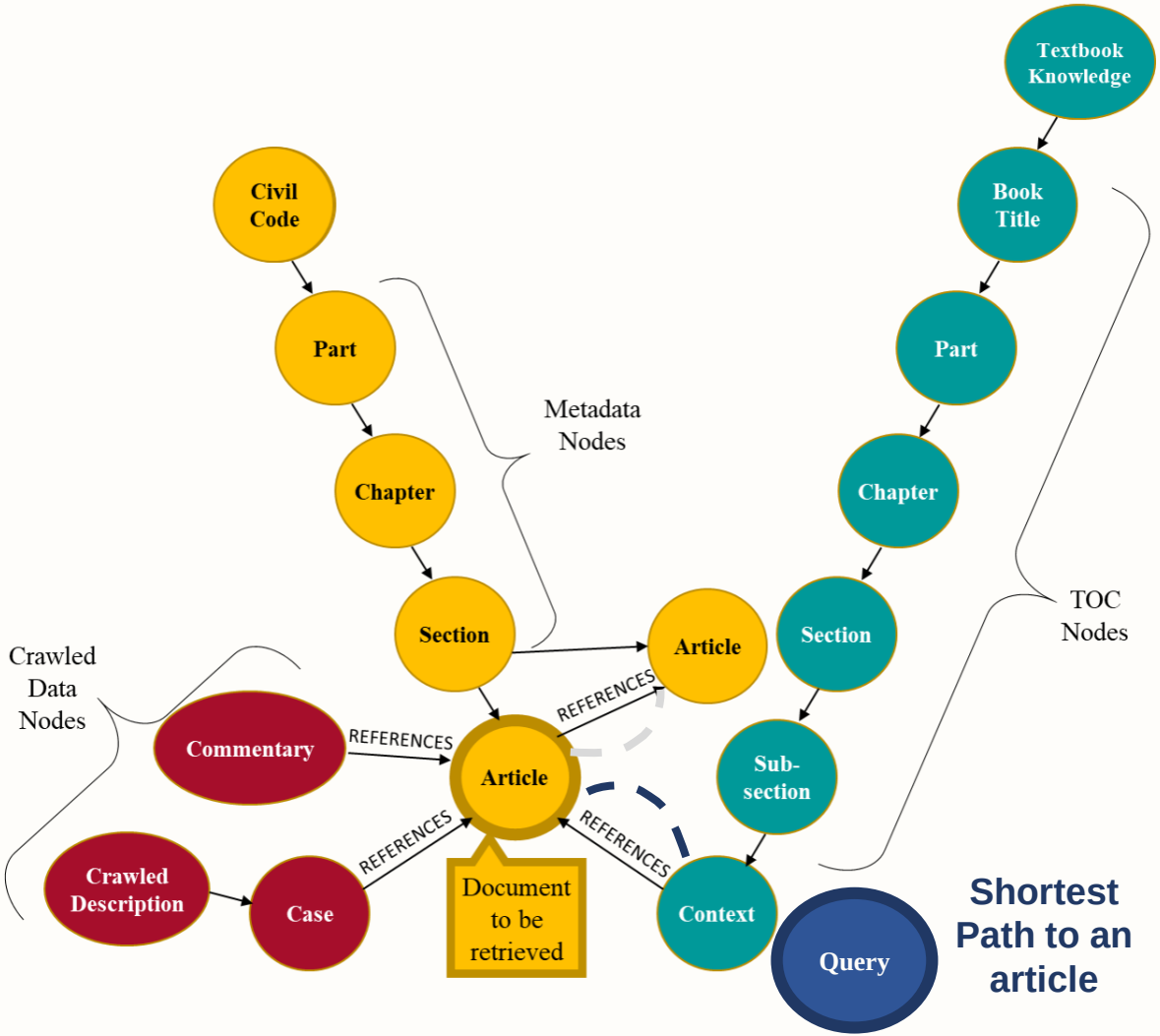
Graph-based Distance



Graph-based Distance



Graph-based Distance



Textual Similarity

Transformer-based Pipeline	Term-based Pipeline
all-mpnet-base-v2	**BM25 with stemming**
bge-m3	BM25 with lemmatization
bge-large-en-v1.5	BM25 with n-grams
e5-large-v2	
all-MiniLM-L6-v2	
all-distilroberta-v1	

Relevance Scoring

hybrid score(article,query) =

w_{art} · **article similarity**
+ w_{comm} · **commentary similarity**
+ w_{prec} · **precedent similarity**
+ w_{cont} · **context similarity**
+ w_{graph} · **hops score**

Textual Similarity

Graph-based Distance

- Between 0,1 (weights add up to 1)

04

Results and Conclusion

Performance vs. Explainability

Best Parameters on Validation Data

BGE-M3

hybrid score(article,query) =

0.8 · article similarity
+ 0.1 · commentary similarity
+ 0.0 · precedent similarity
+ 0.0 · context similarity
+ 0.1 · hops score

Graph Threshold = 0.95
Result Threshold = 0.83

BM25

hybrid score(article,query) =

0.5 · article similarity
+ 0.1 · commentary similarity
+ 0.0 · precedent similarity
+ 0.1 · context similarity
+ 0.3 · hops score

Graph Threshold = 0.8
Result Threshold = 0.89

Results on COLIEE Task 3 Data (2024)

Retrieval Model	Graph Configuration	# Experiments	Best F2 in training data	Best F2 in validation data
BGE-M3 (Finetuned on COLIEE)	Base graph + crawled data + textbook knowledge	5,236	0.837	**0.695**
BGE-M3 (Finetuned on COLIEE)	Base graph + crawled data	1,932	0.837	0.688
BGE-M3 (Finetuned on COLIEE)	Base graph + textbook knowledge	588	0.837	0.690
BM25	Base graph + crawled data + textbook knowledge	5,236	**0.604**	**0.594**
BM25	Base graph + crawled data	1,932	0.601	0.580
BM25	Base graph + textbook knowledge	588	**0.604**	0.579

Conclusion

Trade-off between performance and explainability

Could be observed, but there are always exceptions

Benefit of the graph structure

More queries were answered correctly when the graph was used

Benefit of using the textbooks

Best performances were observed with both, textbook data and further crawled data (commentary, precedent)

Conclusion

Trade-off between performance and explainability

Could be observed, but there are always exceptions

Benefit of the graph structure

More queries were answered correctly when the graph was used

Benefit of using the textbooks

Best performances were observed with both, textbook data and further crawled data (commentary, precedent)

Future Work

Try this approach in a RAG setup

Modern way of introducing justifiability to retrieval results

Thanks!

Do you have any questions?

sabine.wehnert@gei.de

LEIBNIZ INSTITUTE
FOR EDUCATIONAL MEDIA
|Georg Eckert Institute

CREDITS:
Special thanks to the **COLIEE competition** organizers for allowing us to use the Task 3 data.

This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Grid Search Parameters

- w_{art}: [0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]
- w_{comm}: [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]
- w_{prec}: [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]
- w_{cont}: [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7]
- w_{graph}: [0, 0.1, 0.3, 0.5, 0.7, 0.9, 1]

- graph threshold: [0.8, 0.85, 0.9, 0.95]
- result threshold: [0.8, 0.83, 0.86, 0.89, 0.92, 0.95, 0.98]

Category	Coverage for articles	Coverage for articles in %
Crawled data: Commentary	547	70.58%
Crawled data: Precedent	263	33.94%
Crawled data: Reference articles	312	40.26%
Crawled data: Commentary, Precedent	225	29.03%
Crawled data: Commentary, Reference articles	272	35.10%
Crawled data: Precedent, Reference articles	144	18.58%
Crawled data: Commentary, Precedent, Reference articles	133	17.16%
Textbook knowledge	215	27.74%
No Crawled Data	161	20.77%
No Auxiliary Information (total)	149	19.2%

Table 1. Coverage for each category of auxiliary information.

Best Parameters on Training Data

BM25

hybrid score(article,query) =

0.6 · article similarity
+ 0.1 · commentary similarity
+ 0.0 · precedent similarity
+ 0.0 · context similarity
+ 0.3 · hops score

Graph Threshold = 0.9
Result Threshold = 0.8