MUNI
FI

CLEF 2022

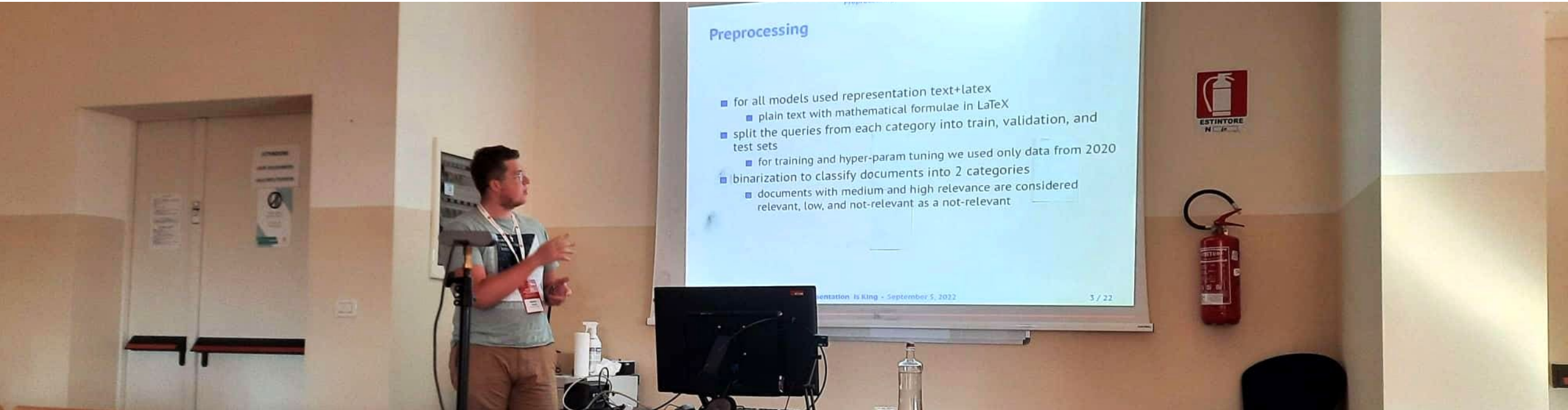
Vít Novotný, Martin Geletka,
Marek Toma, Petr Sojka

About CLEF 2022

- 13th Conference and Labs of the Evaluation Forum [GGS rank B]
- **Venue:** Università di Bologna
- **Topic:** Information access in any modality and language
- **Form:** Workshops presenting results of lab-based benchmarks

FI MU at CLEF 2022

- **Vítek Novotný** co-organized the ARQMath-3 lab about MathIR
- **Martin Geletka** and **Marek Toma** presented the best automatic run at ARQMath-3 Task 1 (Answer Retrieval) among 7 teams.




SNAP4

ARQMath-3

Answer Retrieval for Questions on Math

https://www.cs.rit.edu/~dprl/ARQMath

#ARQMath

Behrooz Mansouri
Rochester Institute of Technology USA
bm3302@rit.edu

Vítek Novotný
Faculty of Informatics, Masaryk University, Czechia
witiko@mail.muni.cz

Anurag Agarwal
Rochester Institute of Technology USA
axasma@rit.edu

Douglas W. Oard
University of Maryland USA
oard@umd.edu

Richard Zanibbi
Rochester Institute of Technology USA
rxzvcs@rit.edu

Task 1: Find Answers to Math Questions

Given one of Math Stack Exchange question as a query,
search answer posts from prior years, **return up to 1,000 answers**

Question about determinant of a block matrix

I was studying block matrices and suddenly this question came to my mind.

Let $A, B \in \mathbb{R}^{n \times n}$. From this Wikipedia page,

$$\det \begin{pmatrix} A & B \\ B & A \end{pmatrix} = \det(A - B)\det(A + B)$$

even if A and B do not commute. Does a similar condition hold for the following block matrix?

$$\begin{pmatrix} A & -B \\ B & A \end{pmatrix}$$

matrices determinant block-matrices

Topic A.348

Look at the arithmetic operations and their actions. With + and *, these matrices form a field. And we have the isomorphism

$$a + ib \mapsto \begin{bmatrix} a & -b \\ b & a \end{bmatrix}.$$

Assessed as Medium Relevance

As Lord Farin points out, this is certainly false. Consider for example any prime p and the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

which satisfies $A^x = 0$ for all $x > 1$.

Assessed as Non-Relevant

Task 2: Contextualized Formula Retrieval

Given a **formula** from a Task 1 question,
search questions & answers from prior years, return relevant formulae

Question about determinant of a block matrix

0 I was studying block matrices and suddenly this question came to my mind.

Let $A, B \in \mathbb{R}^{n\times n}$. From this Wikipedia page,

$$\det\begin{pmatrix} A & B \\ B & A \end{pmatrix} = \det(A-B)\det(A+B)$$

even if A and B do not commute. Does a similar condition hold for the following block matrix?

$$\boxed{\begin{pmatrix} A & -B \\ B & A \end{pmatrix}}$$

matrices | determinant | block-matrices

Topic B.348

There are regular graphs that are not distance-regular but do have perfect 1-codes. If A and B are the adjacency matrices of a graph X and its complement, the matrix

$$\boxed{\begin{pmatrix} A & B \\ B & A \end{pmatrix}}$$

is the adjacency matrix of a graph with a perfect 1-code of size two (in most cases). (There are many other examples, but these are the first that come to mind.)

Assessed as Medium Relevance

How to prove this trace matrix inequality?

4 Given:

1. A is a diagonal positive definite matrix;
2. $\mathrm{Tr}(A) = 1$ and $\mathrm{Tr}(A^2) \le 1$;
3. B is a Hermitian matrix;
4. $\boxed{AB \neq BA}$.

How to prove the following:

$$\mathrm{Tr}(AABB) - \mathrm{Tr}(ABAB) \le \mathrm{Tr}(A^2)\,[\mathrm{Tr}(ABB) - \mathrm{Tr}(AB)\,\mathrm{Tr}(AB)]?$$

Assessed as Non-Relevant

Task 3: Open-Domain Question Answering

Given a Math Stack Exchange question as a query (as Task 1), return **a single (extracted/generated)** answer to math questions

Number of solutions of equation over a finite field

Asked 1 year ago Modified 1 year ago Viewed 54 times

1

I have a question regarding the number of solutions of a equation over a finite field $\mathbb{F}_p$. First of all, consider the equation $x^3 = a$ over $\mathbb{F}_p$, where p is a prime such that $p \equiv 2 (\mathrm{mod}\ 3)$. The book that I'm currently reading says that this equation has exactly one solution in $\mathbb{F}_p$ for every $a \in \mathbb{F}_p$, because $\gcd(3, p-1) = 1$, but the book does not prove this. Unfortunately, this doesn't convince me enough. Is there is a convincing elementary straightforward proof justifying why is this true?

number-theory elementary-number-theory finite-fields

Share Cite Edit Follow Flag

edited Jul 30, 2021 at 4:05
davidlowryduda
86.9k 9 157 299

asked Jul 30, 2021 at 0:03
user463019
387 1 7

Topic A.309

If $a \neq 0$, then $x^3 - a = (x-r)(x^2 + rs + s^2)$. Because $\gcd(3, p-1) = 1$, there are units r and s such that $r + s = -1$. This implies that $x^3 - a$ factors into linear terms, so it is not separable, so it has a repeated root r and therefore has at most one solution. Now, suppose $a = 0$. Then $x^3 = 0$ has at most one solution because $\gcd(3, p) = 1$.

Assessed as Highly Relevant

Hint: consider $\sum_{t=0}^{p-1} \sum_{x_i} \sum_{x_2} \cdots \sum_{x_n} e^{2\pi i t(x_1^2 + x_2^2 + \cdots + x_n^2)/p}$
The sums on the x_i all run from 0 to $p-1$.

Assessed as Non-Relevant

Task 3: Runs

Runs

- 1 baseline run:
  - **GPT-3** (automatic, generative)
- 13 participant runs from 3 teams:
  - 5 runs from **Approach0** (manual, extractive)
  - 4 runs from **DPRL** (automatic, extractive)
  - 4 runs from **TU_DBS** (automatic, generative)

Brown et al., Language Models are Few-Shot Learners. NeurIPS, 2020

Task 3 Baseline Run: GPT-3

We use text-davinci-002 model of GPT-3 from OpenAI

First, we prompt GPT-3 as follows:

```
Q: What does it mean for a matrix to be Hermitian?

A:
```

Brown et al., Language Models are Few-Shot Learners. NeurIPS, 2020

Task 3 Baseline Run: GPT-3

We use text-davinci-002 model of GPT-3 from OpenAI

GPT-3 completes the text and produces an answer:

```
Q: What does it mean for a matrix to be Hermitian?

A: A matrix is Hermitian if it is equal to its
transpose conjugate.
```

Brown et al., Language Models are Few-Shot Learners. NeurIPS, 2020

Task 3: Evaluation

Manual Evaluation Measures

- Average Relevance (AR)
- Precision at 1 (P@1)

Automatic Evaluation Measures

- Lexical Overlap (LO)
- Contextual Similarity (CS)

Task 3 Evaluation: Manual Measures

Average Relevance (AR)

A.301 Not Relevant (0)	A.302 Medium Relevance (2)	A.303 Low Relevance (1)

AR = (0 + 2 + 1) / 3 = 1.00

Precision at 1 (P@1)

A.301 Not Relevant (0)	A.302 Medium Relevance (2)	A.303 Low Relevance (1)

P@1 = (0 + 1 + 0) / 3 = 0.33

Task 3 Evaluation: Automatic Measures

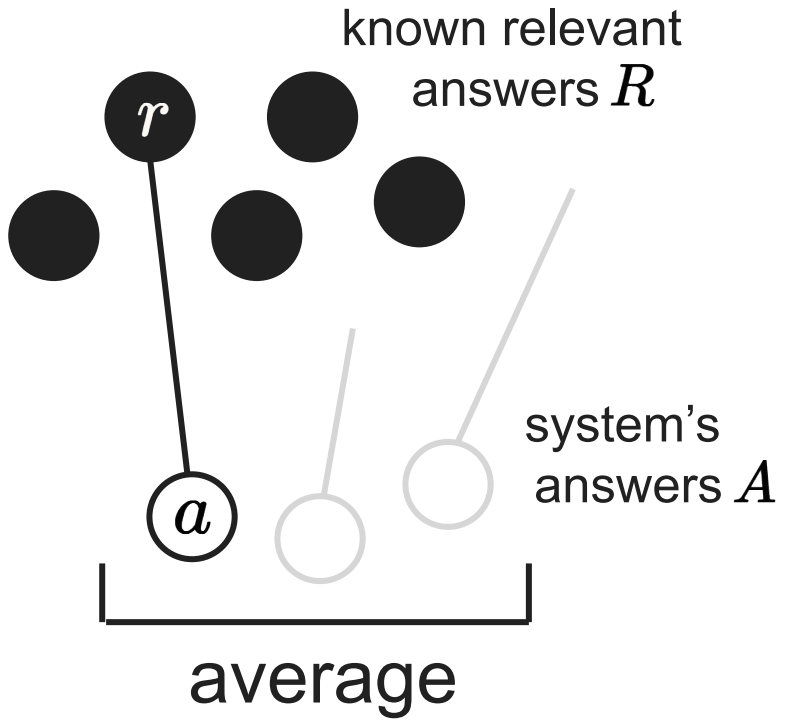
Lexical Overlap (LO) and Contextual Similarity (CS)

$\frac{1}{|A|} \cdot \sum_{a \in A} \max_{r \in R} \text{similarity}(a, r)$, where A are the system's answers to a question and R are known relevant answers for the same question

Question:
What does it mean for a matrix to be Hermitian?

System's Answer a:
A matrix that is equal to its transpose conjugate

Known Relevant Answer r:
A complex square matrix that is equal to its own conjugate transpose



Task 3 Evaluation: Automatic Measures

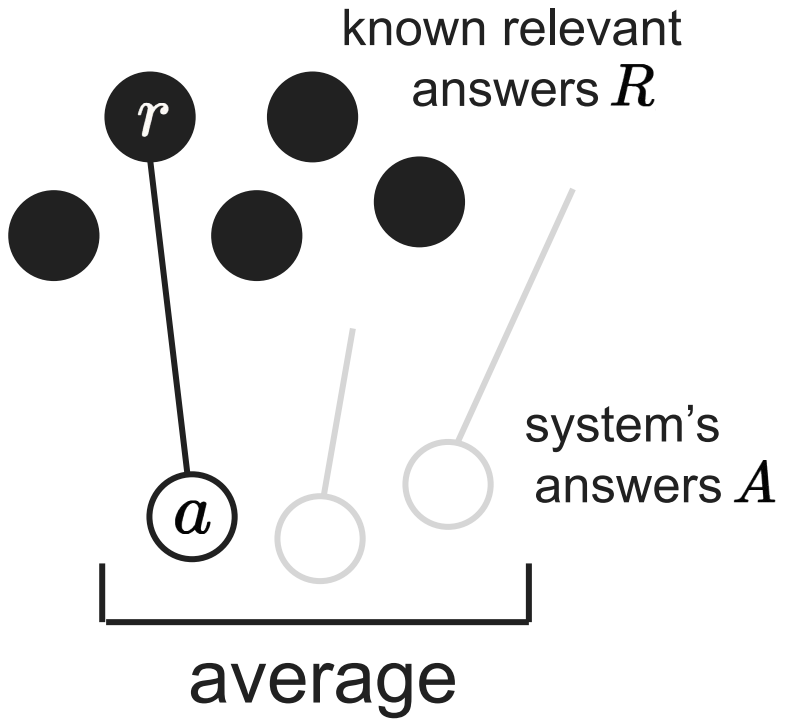
Lexical Overlap (LO)

$\frac{1}{|A|} \cdot \sum_{a \in A} \max_{r \in R} F_1\text{-score}(a, r)$, where A are the system's answers to a question and R are known relevant answers for the same question

Contextual Similarity (CS)

$\frac{1}{|A|} \cdot \sum_{a \in A} \text{BERTScore}(a, R)$

(a, r)



T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT. ICLR 2020

V. Novotný and M. Štefánik. Combining Sparse and Dense Information Retrieval. Soft Vector Space Model and MathBERTa at ARQMath-3 Task 1 (Answer Retrieval). CLEF 2022

Task 3 Results: Best Run per Team

			Run Type			ARQMath-3 (78 Topics)				(73 Topics)	
Team	Run	Data	Primary	Manual	Generative	AR	P@1	LO	CS	MG	UI
Baseline	GPT-3	Both	✓			**(1.346)**	**(0.500)**	0.317	0.851	**0.288**	**(0.466)**
Approach0	run1	Both		✓		**1.282**	**0.436**	**0.509**	**0.886**	**0.110**	**0.562**
DPRL	SBERT-SVMRank	Both				0.462	0.154	0.330	0.846	0.205	0.767
TU_DBS	amps3_se1_hints	Both				0.325	0.078	0.263	0.835	0.833	0.931

Manual Evaluation

- **GPT-3** outperformed all runs; **Approach0** run is a close second

Task 3 Results: Best Run per Team

			Run Type			ARQMath-3 (78 Topics)				(73 Topics)	
Team	**Run**	**Data**	Primary	Manual	Generative	AR	P@1	LO	CS	MG	UI
Baseline	GPT-3	Both	✓		🤖	**(1.346)**	**(0.500)**	0.317	0.851	**0.288**	**(0.466)**
Approach0	run1	Both		✓	👤	**1.282**	**0.436**	**0.509**	**0.886**	**0.110**	**0.562**
DPRL	SBERT-SVMRank	Both			👤	0.462	0.154	0.330	0.846	0.205	0.767
TU_DBS	amps3_se1_hints	Both			🤖	0.325	0.078	0.263	0.835	0.833	0.931

Manual Evaluation

- **GPT-3** outperformed all runs; **Approach0** run is a close second

Automatic Evaluation

- **Lexical Overlap (LO)** correlates with manual measures (τ = 0.736)
- LO can be used to evaluate future systems for Open Domain QA

	AR	**P@1**	**LO**	**CS**
AR	1.000	0.994	0.736	0.670
P@1		1.000	0.729	0.674
LO			1.000	0.805
CS				1.000

Kendall's τ

Task 3 Post-Evaluation: Characterizing Answers

In addition to quantitative evaluation, we were interested in the following:

- *Can assessors distinguish human and machine-generated answers?*
- *Do Task 3 systems stuff answers with unrelated information?*

Task 3 Post-Evaluation: Characterizing Answers

In addition to quantitative evaluation, we were interested in the following:

- *Can assessors distinguish human and machine-generated answers?*
- *Do Task 3 systems stuff answers with unrelated information?*

We provided a sample of Task 1 and Task 3 answers to assessors, and asked:

- Whether they thought the answers were machine-generated
- Whether answers contained information unrelated to the topic question

Task 3 Post-Evaluation: Characterizing Answers

In addition to quantitative evaluation, we were interested in the following:

- *Can assessors distinguish human and machine-generated answers?*
- *Do Task 3 systems stuff answers with unrelated information?*

We provided a sample of Task 1 and Task 3 answers to assessors, and asked:

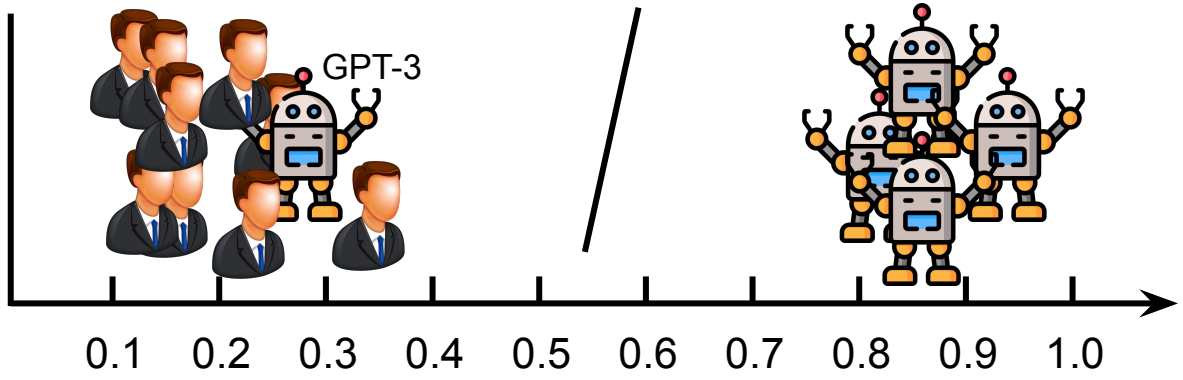
- Whether they thought the answers were machine-generated
- Whether answers contained information unrelated to the topic question

We report the following post-evaluation measures:

- Machine-Generated (MG) – Fraction of answers assessed as machine-generated
- Unrelated Information (UI) – Fraction of answers with unrelated information

Task 3 Results: Best Run per Team

			Run Type			ARQMath-3 (78 Topics)				(73 Topics)	
Team	Run	Data	Primary	Manual	Generative	AR	P@1	LO	CS	MG	UI
Baseline	GPT-3	Both	✓			**(1.346)**	**(0.500)**	0.317	0.851	**0.288**	**(0.466)**
Approach0	run1	Both		✓		**1.282**	**0.436**	**0.509**	**0.886**	**0.110**	**0.562**
DPRL	SBERT-SVMRank	Both				0.462	0.154	0.330	0.846	0.205	0.767
TU_DBS	amps3_se1_hints	Both				0.325	0.078	0.263	0.835	0.833	0.931



Separation of extractive ( ) / generative ( ) runs using **MG**

Characterizing Answers

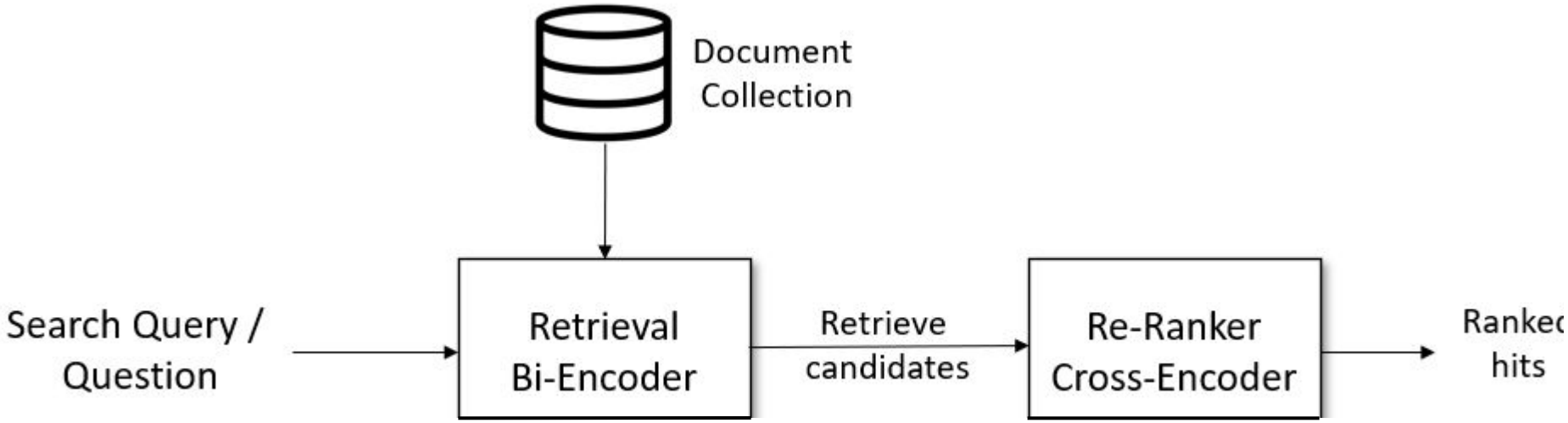
- Assessors reliably identified machine-generated answers with the exception of GPT-3 (MG = 0.288).
- Anti-correlation between effectiveness and unrelated information (τ = –0.88) indicates no answer stuffing.


SALA RUSCONI
10:17 300km/h 18

MIR systems on ArqMath 2022

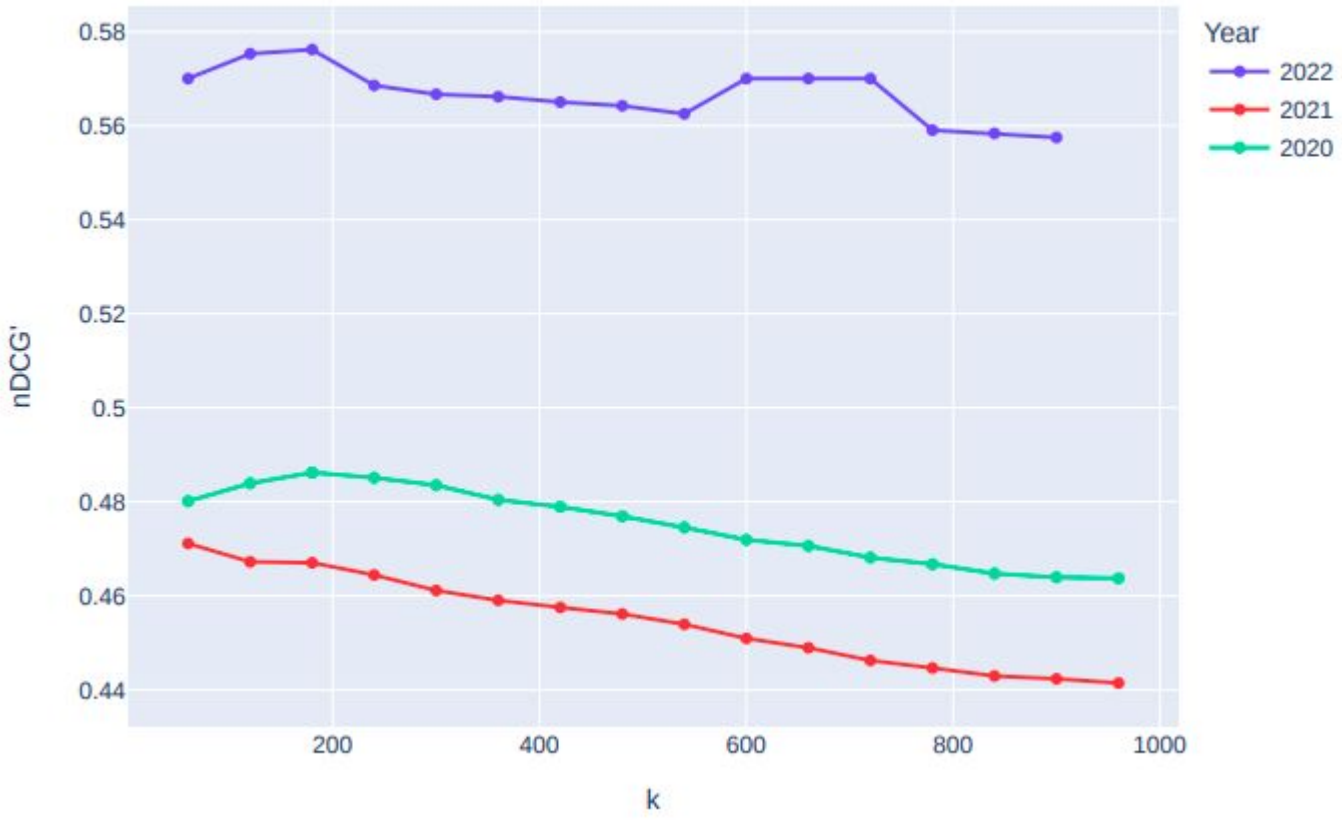
- Runs from PV211 students (MSM team)
  - TF-IDF, BM25, CompuBERT
- Runs submitted by MIR teams
  - Variations of deep Retrieval / ReRanker models
- Ensembles of of individual systems
  - IBC, RRF, RBC, WIBC

Retrieval - ReRanker models



MUNI
FI

Best system - RRF ensemble



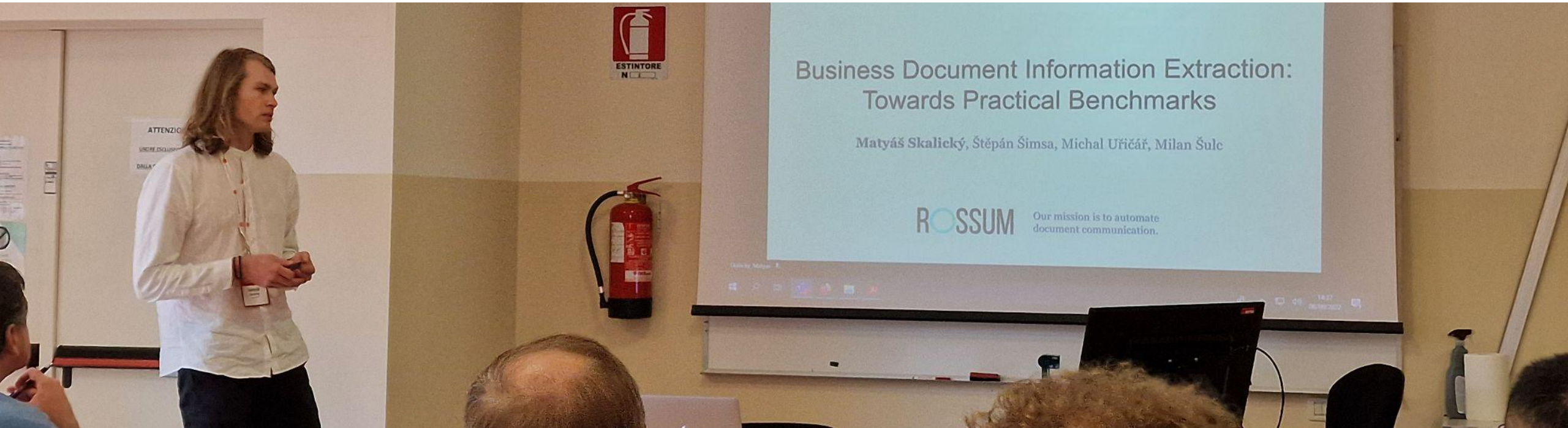
MUNI FI

Rossum.ai at CLEF 2022

- Need for practical Benchmark
  - Business Document Information Extraction: Towards Practical Benchmarks
- Article describes:
  - Need for practical benchmark for DocILE
  - Main problems researched by Rossum.ai
    - Key Information Extraction and Localization
    - Table Extraction and Line Items
    - One-Shot Learning for Information Extraction
  - Other related problems:
    - Optical Character Recognition
    - Document Layout Analysis
    - Extraction of Key-Value Pairs
    - Question Answering

DocILE at CLEF 2023

- Document Information Localization and Extraction
- Lab proposal by ROSSUM, to be held bi-yearly at CLEF
- **Goal:** Industry-strength benchmarks for invoice-like documents



Other visited labs on CLEF

- Image CLEF
  - One of the biggest labs on CLEF conference
  - ImageCLEFaware, ImageCLEFcoral, ImageCLEFmedical, ImageCLEFfusion
  - SnakeCLEF - organized by University of West Bohemia

- CheMu
  - Task 1 Expression level extraction
    - Named Entity Recognition
    - Event Extraction
    - Anaphora Resolution
  - Task 2 Document level information extraction
    - Chemical Reaction Reference Resolution
    - Table Semantic Classification

Other visited labs on CLEF

- eRisk
  - Early risk prediction on Internet
  - Early Detection of Signs of Pathological Gambling
  - Early Detection of Depression
  - Measuring the severity of the signs of Eating Disorders

- Check That
  - Fighting the Covid-19 Disinformation and Fake news Detection
  - Identification of Relevant Claims on Twitter
  - Detecting Previously Fact-Checked Claims
  - Fake News Detection

MathIRQA at ECIR 2023

- Math-aware Information Retrieval and Question Answering
- Workshop proposal by FI MU and NIT Silchar (India)
- Topics and Themes:
  - Math information retrieval
  - Representation of math information
  - Formula Search
  - Math-aware question answering
  - Math problem solving
  - Semantic interpretation of math information
  - Index optimization
  - Scientific document retrieval
  - Scientific information extraction
  - Discovery of scientific knowledge
  - Searching & ranking of math information
  - Formula embedding

MUNI FI