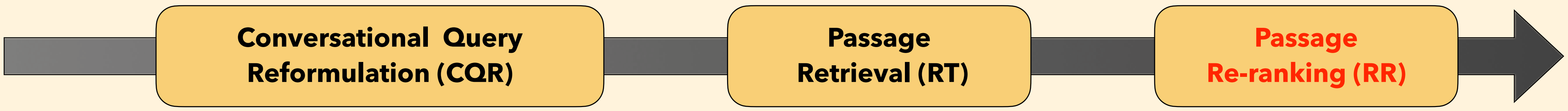
Improving Conversational Passage Re-ranking via View Ensemble

Jia-Huei Ju[1], Sheng-Chieh Lin[2], Ming-Feng Tsai[3], and Chuan-Ju Wang[1]
[1]Academia Sinica, [2]University of Waterloo, [3]National Chengchi University

Conversational Search – The Multi-stage Pipeline

Conversational query (vs. ad-hoc query)

✓ contains an *user utterance* (e.g. a question) and *conversational context* (e.g., previous asked questions)



$$q'_i = \mathscr{F}_{CQR}(u_i; u_1, u_2, \ldots, u_{i-1}) \qquad \mathscr{D}_{\mathrm{RT}} = \mathscr{F}_{RT}(q'; p \in \mathscr{D}) \qquad R = \mathscr{F}_{\mathrm{RR}}(q'; p \in \mathscr{D}_{\mathrm{RT}})$$

Conversational dense retrieval (ConvDR)

✓ Integrates CQR into dense retrieval models by retrofitting the *query encoder* (e.g., ConvDR, CQE, …etc.)

Research Question: Can we have a better passage re-ranker (ConvRerank) for the pipeline?

✓ Effectiveness: **Tok-ranking** sensitive (e.g., nDCG@3)

✓ Efficiency: (i) Discards the CQR module; (ii) performs re-rank on **limited** passages (e.g., top-100).

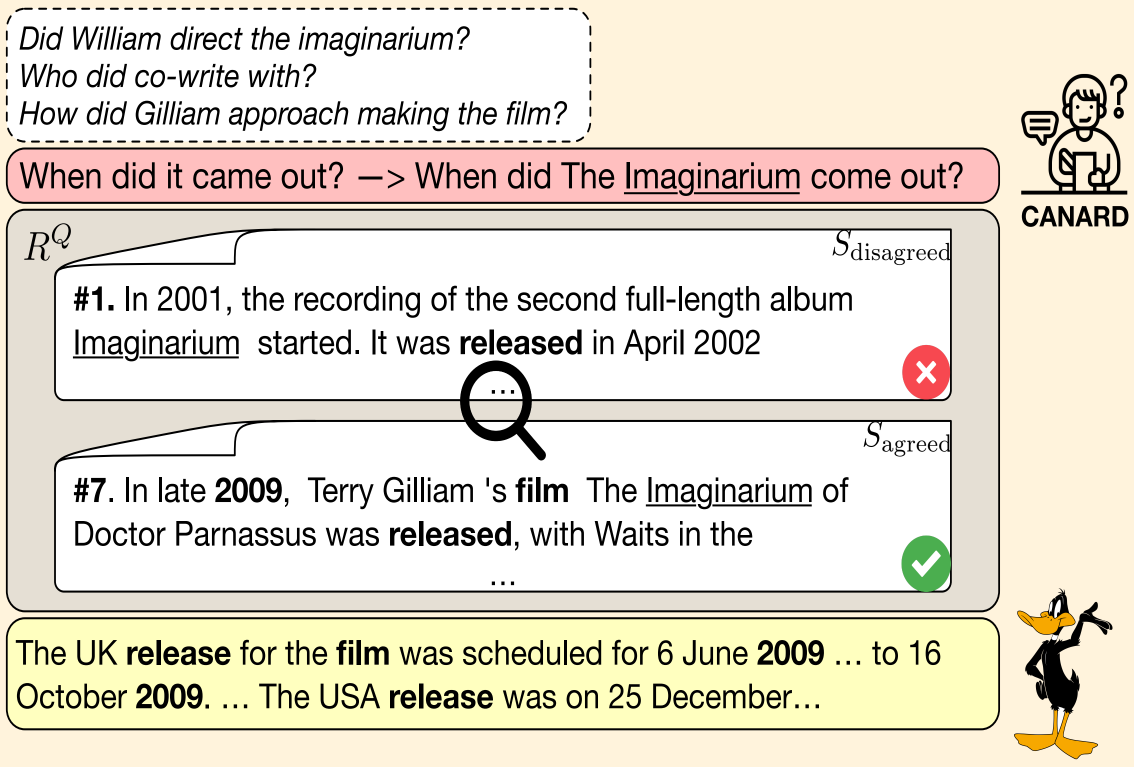
Methods – ConvRerank Fine-tuned on Pseudo-labeled Dataset with View Ensemble

Motivation

✓ Pseudo-labels (i.e., passage relevances) sometimes **conflicts** with corresponding conversational *context.*

✓ **Ground-truth answers** should be able to **calibrate**.

Our goal: Ensemble the relevance of Question and Answer view by mixing two ranked lists.



Did William direct the imaginarium?
Who did co-write with?
How did Gilliam approach making the film?

When did it came out? —> When did The Imaginarium come out?

R^Q

$S_{\text{disagreed}}$

#1. In 2001, the recording of the second full-length album Imaginarium started. It was **released** in April 2002

…

S_{agreed}

#7. In late **2009**, Terry Gilliam 's **film** The Imaginarium of Doctor Parnassus was **released**, with Waits in the …

The UK **release** for the **film** was scheduled for 6 June **2009** … to 16 October **2009**. … The USA **release** was on 25 December…

Procedures

✓ First, we construct an initial ranked list

✓ Second, we concatenate question with the answer for constructing ranked list R^A as an another view.

✓ Finally, pushing passages both appeared (agreed) in two lists to the top; and the other to the bottom.

$$R^A = \mathrm{monoT5}\big(q^*; p \in \mathrm{BM25}(q^* \parallel a; p \in \mathcal{D})\big),$$

$$R^Q = \mathrm{monoT5}\big(q^*; p \in \mathrm{BM25}(q^*; p \in \mathcal{D})\big),$$

$$R^{\mathrm{EM}(R^Q|R^A)} = \Phi(R^Q, R^A) = S_{\text{agreed}} \parallel S_{\text{disagreed}},$$

✓ Then, fine-tune monoT5 on this data as ConvRerank.

Evaluation – TREC CAsT 2019 & 2020

	Latency	CAsT'19 Eval	CAsT'20 Eval
Retrieval (→ Re-ranking)	(ms/q)	nDCG@3 / 100	nDCG@3 / 100
Upper-bound system w/ manual query			
TCT-ColBERT [19] → monoT5	-	0.583 / 0.545	0.556 / 0.546
ConvDR → BERT (RRF) [40]	1900	0.541 / -	0.392 / -
CRDR [26]	1690	0.553 / -	0.381 / -
CTS+MVR† [15]	14630	**0.565** / -	- / -
CQE	-	0.492 / 0.447	0.319 / 0.350
CQE → T5-rewrite+monoT5	1910	0.549^d / 0.484^d	0.418^d / 0.395^d
CQE → ConvRerank	1675	$\mathbf{0.563}^d$ / $\mathbf{0.487}^d$	$\mathbf{0.432}^d$ / $\mathbf{0.456}^{de}$

Comparison with Different Pseudo-labels

	CAsT'19 Eval	CAsT'20 Eval
Ranked list	nDCG@3 / 100	nDCG@3 / 100
$R^{\mathrm{EM}(R^Q\|R^A)}$ (proposed)	$\mathbf{0.563}^{bcd}$ / $\mathbf{0.487}^{bcd}$	$\mathbf{0.432}^{bcd}$ / $\mathbf{0.456}^{bcd}$
R^Q	0.517 / 0.467	0.396 / 0.382
R^A	0.495 / 0.464	0.392 / 0.382
$R^{\mathrm{EM}(R^A\|R^Q)}$	0.519^c / 0.474^{bc}	0.403 / 0.389^{bc}