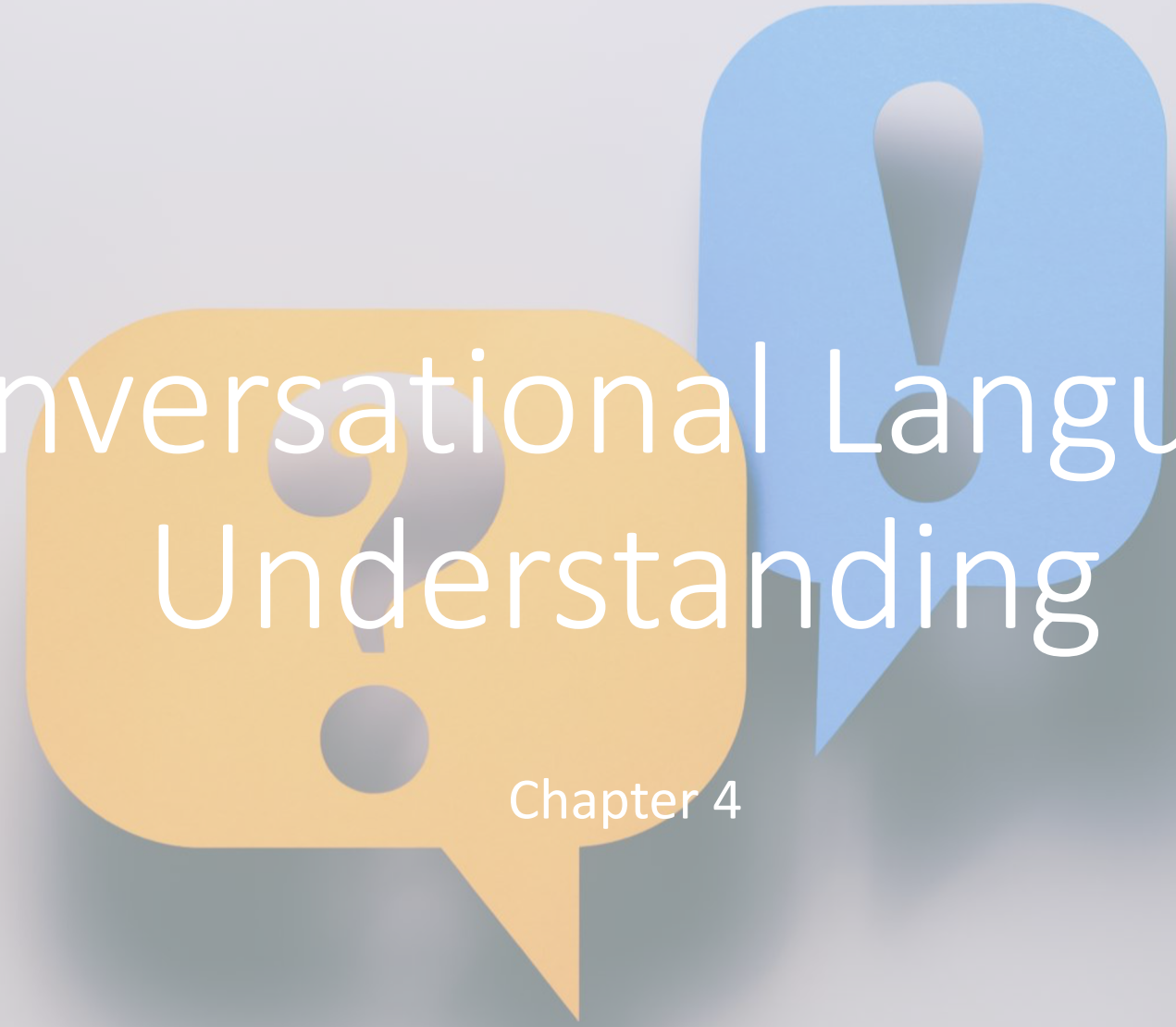


Conversational Language Understanding

Chapter 4

A reminder on task terminology & datasets

- Conversational Question Answering (**ConvQA**)
  SequentialQA, **QuAC**, **CoQA**, QReCC, TopioCQA (new!)
- Knowledge graph Conversational Question Answering (**KG-ConvQA**)
  Complex Sequential QA, ConvQuestions
- Conversational Passage Retrieval (**ConvPR**)
  TREC CAsT, QReCC
- Conversational Document Retrieval (**ConvDR**)
  TREC CAsT

Conversation modeling

Example CAsT Y3 topic: genetic engineering

- How do genes work?
- What other diseases are caused by a single change?
- What are the other types of diseases?
- You missed the second type of disease. What was that?
- That's not what I wanted. How about recent developments in gene therapy to treat those defects?

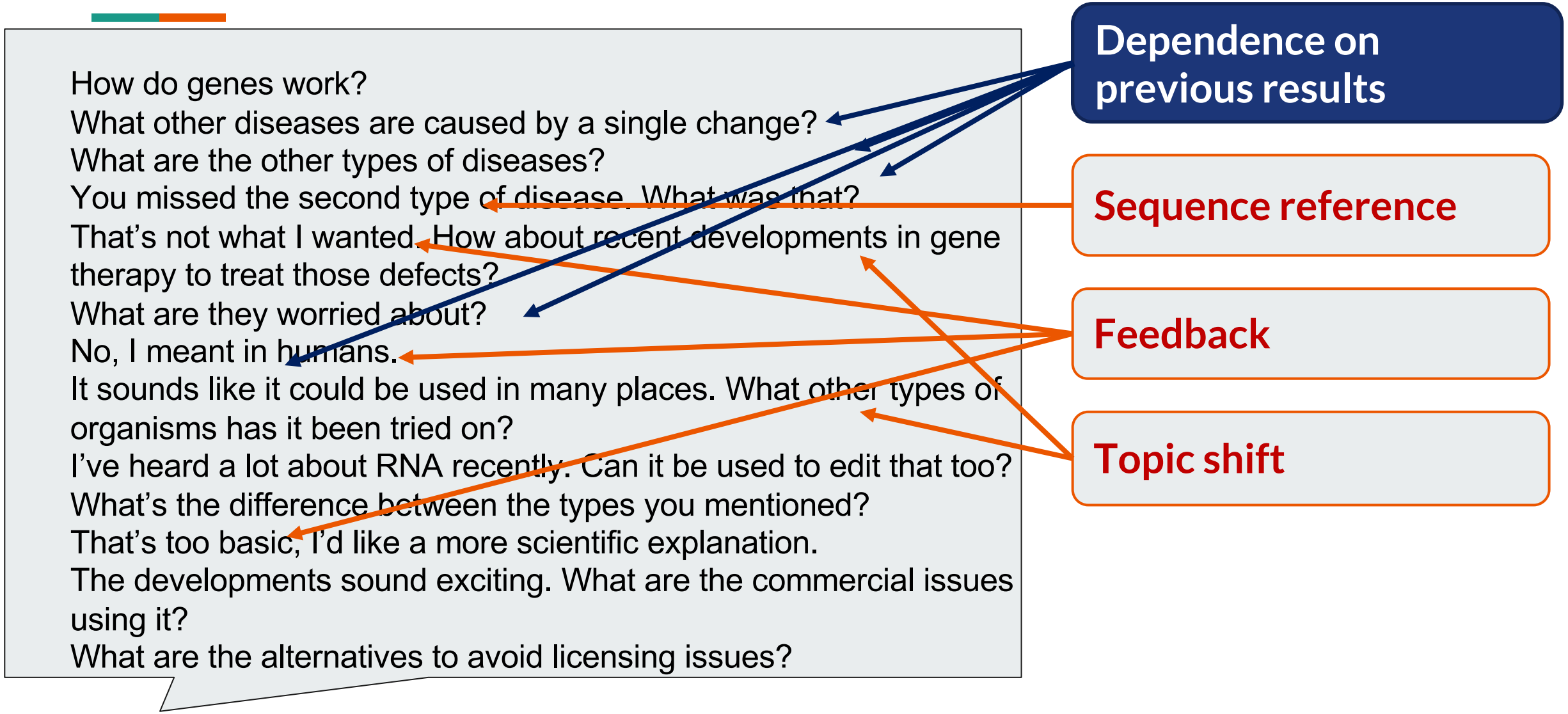
A large genetic mistake typically occurs in the woman's egg, which may partially explain why older women are more likely to have babies with Down syndrome… Down syndrome is the most common and well-known chromosome defect, but there are many. Types of chromosome diseases: There are several common types of chromosome errors that cause disease. The effects of errors in the sex chromosomes (X and Y) differ greatly from errors in the autosomes (chromosomes 1..22). ***(*MARCO_D76761*)***

Y3 topic: genetic engineering

- How do genes work?
- What other diseases are caused by a single change?
- What are the other types of diseases?
- You missed the second type of disease. What was that?
- That's not what I wanted. How about recent developments in gene therapy to treat those defects?
- What are they worried about?
- No, I meant in humans.
- It sounds like it could be used in many places. What other types of organisms has it been tried on?
- I've heard a lot about RNA recently. Can it be used to edit that too?
- What's the difference between the types you mentioned?
- That's too basic, I'd like a more scientific explanation.
- The developments sound exciting. What are the commercial issues using it?
- What are the alternatives to avoid licensing issues?

Y3 topic: genetic engineering

How do genes work?
What other diseases are caused by a single change?
What are the other types of diseases?
You missed the second type of disease. What was that?
That's not what I wanted. How about recent developments in gene therapy to treat those defects?
What are they worried about?
No, I meant in humans.
It sounds like it could be used in many places. What other types of organisms has it been tried on?
I've heard a lot about RNA recently. Can it be used to edit that too?
What's the difference between the types you mentioned?
That's too basic, I'd like a more scientific explanation.
The developments sound exciting. What are the commercial issues using it?
What are the alternatives to avoid licensing issues?



Widely varying discourse structure

QuAC

- Continuation
  - (follow up, maybe follow up, or don't follow up)
- Affirmation
  - (yes, no, or neither)
- Answerability
  - (answerable or no answer)

TREC CAsT (Y3)

- Questions (~85%),
- Feedback (10%),
- Revealment (5%),
- Elaboration (5%)

Most existing CIS datasets and models have limited discourse types with **users asking questions**, with the **system responding** with **answers** or a clarifying question.

CAsT Y3 discourse examples

Feedback ~ 10% of turns

Revealment ~5% of turns

Elaboration ~5% of turns

- Does the article have more about it?
- Could you expand on some of these methods?
- Give me some examples.
- Tell me more about them.

four weeks.

[Dalton et al. 2021]

Conceptualizing Agent-Human Interactions

- A taxonomy of User and Agent CIS behaviors.

User			Agent		
	Query Formulation	**Reveal**	**Inquire**	User Revealment	Memory
		Disclose, Non-Disclose	Extract, Elicit, Clarify		
		Revise, Expand	Elaborate		
Set Retrieval	Result Exploration	**Inquire**	**Reveal**	System Revealment	
		List, Summarize, Compare	List, Summarize, Compare		
		Subset, Similar	Subset, Similar		
		Navigate	**Traverse**		
		Repeat, Back, More,…, Note	Repeat, Back, More,…, Record		
Mixed Initiative					
		Interrupt	**Suggest**		
		Question, Stop, Change	Recommend, Hypothesize		
			Voice Opinion		
		Interrogate	**Explain**		
		Understand, Explain	Report, Reason		

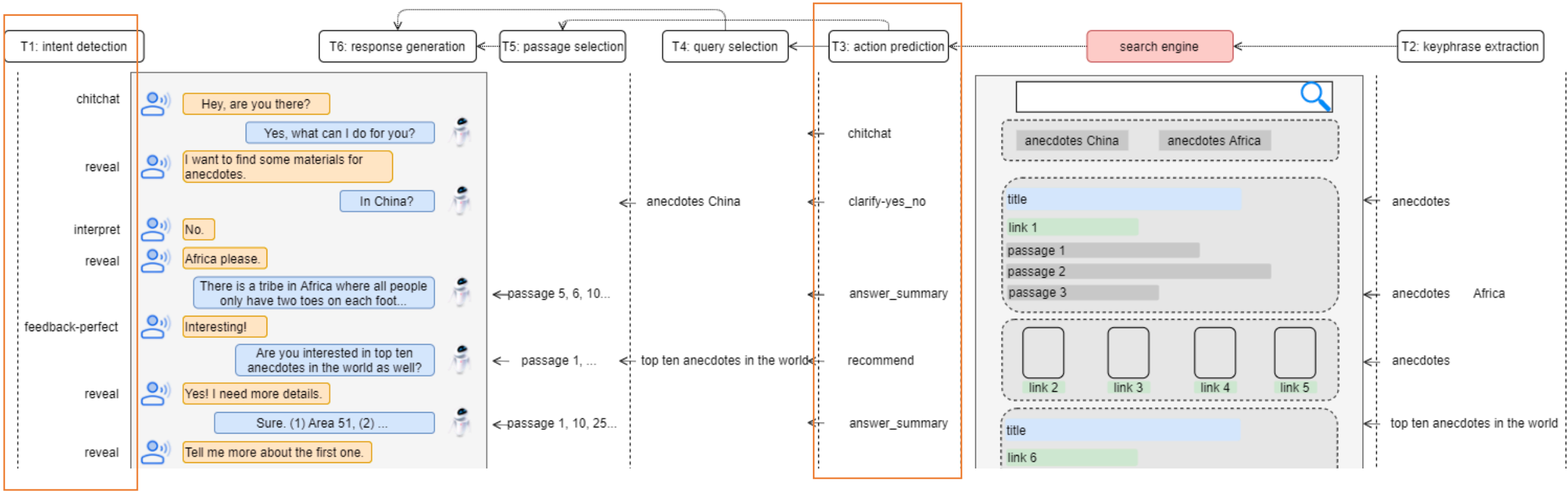
User: I would to arrange a holiday to Italy [Disclose - Volunteer]
Agent: When would to go on holidays?
User: The 4th of May [Disclose - Inquire].
…
Agent: Do you where in Italy you like to go on holidays?
User: I'm not sure [Disclose - Unsure].
…
Agent: What is your budget?
User: I'd prefer not to say [Disclose - Not].

User: Tell me about all the different things you can do in Tuscany? [Inquire List]
…
User: Can you give me an overview of the things to do there? [Inquire Summarize]
…
User: What is the best thing to do in Tuscany? [Inquire Subset]
…
User: Which things are suitable for children? [Inquire Subset]
…
User: What are the main differences between Tuscany and Galicia? [Inquire Compare]
…
User: What other regions in Europe are like that? [Inquire Similar]

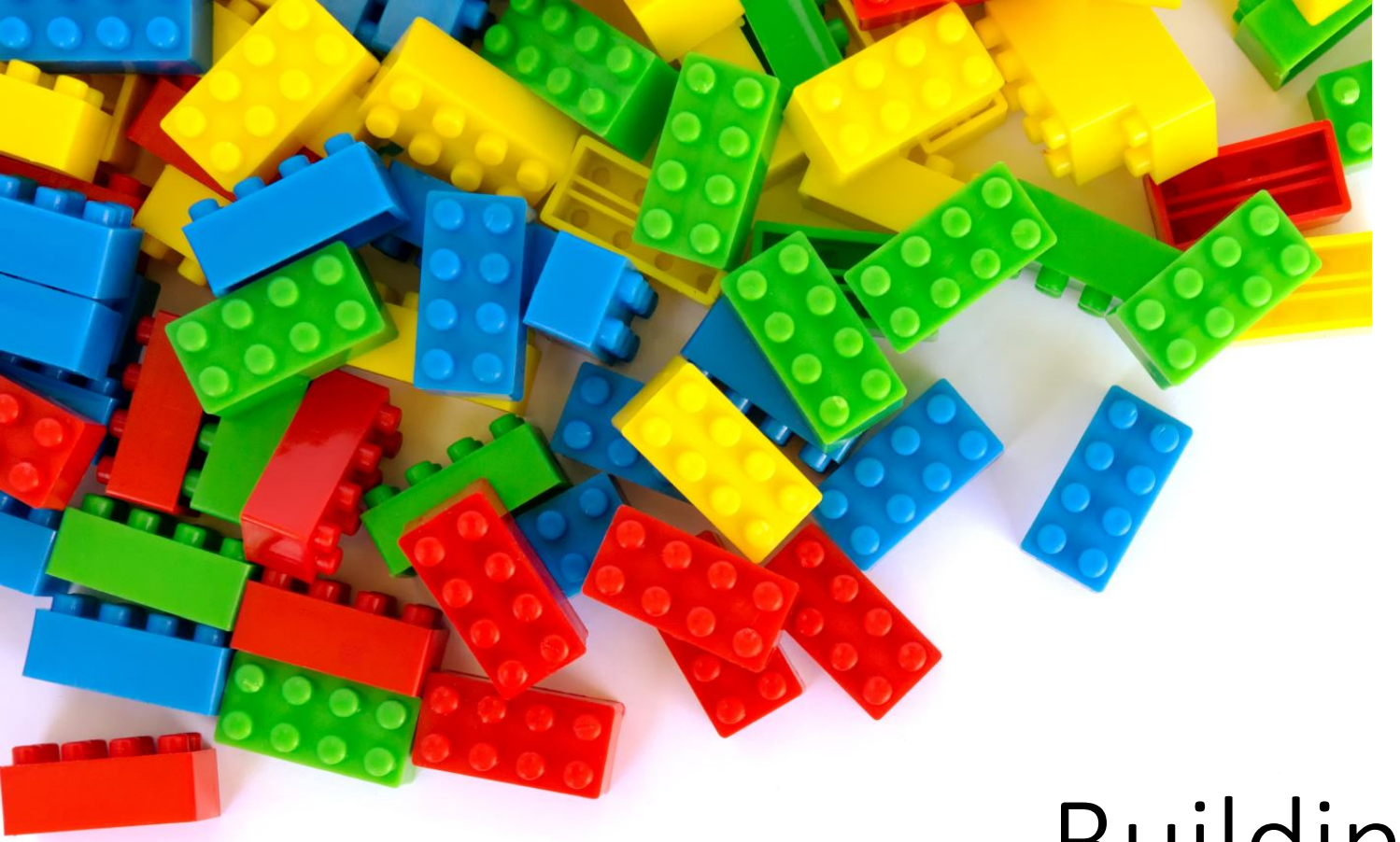
[Azzopardi et al., 2018]

Wizard of Search-Engine

- User intents: reveal, revise, interpret, request-rephrase, chitchat
- Intermediary: clarify, answer-type, answer-form, no-answer, request-rephrase, chitchat



[Ren et al., 2021]



Building blocks

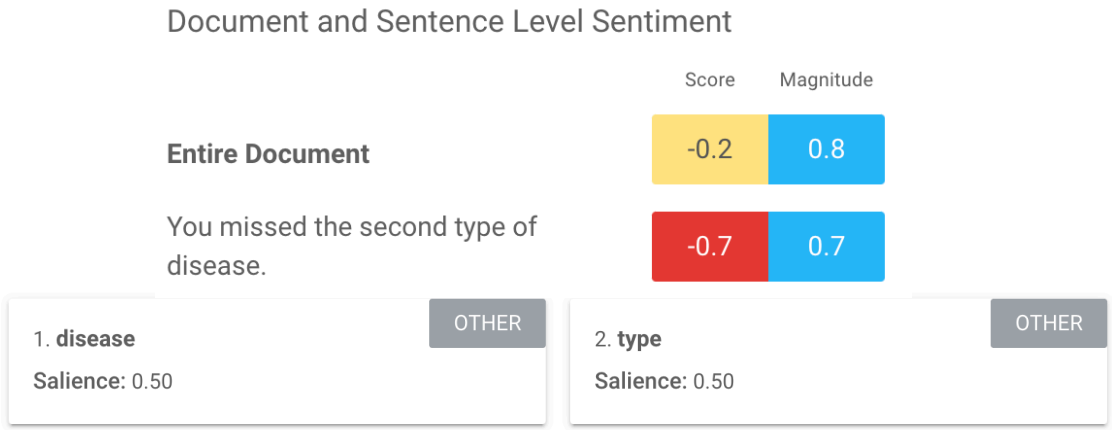
What is CIS turn state?

Representing a single utterance

- Words

[You, missed, the, second, type, of, disease,. What, was, that,?]

- Standard NLP annotations
  - Sentiment
  - Entities
  - …



- Discourse/intent classification(s)

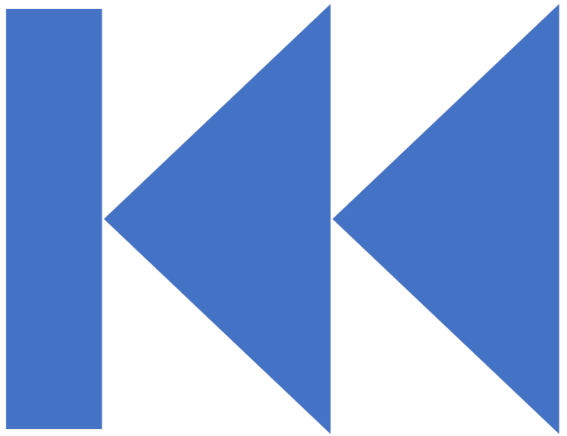
[You missed the second type of disease. What was that?] – **Intent**: **Clarification / Feedback**

- Embedding at token or sequence level
  - BERT, S-BERT, ANCE, etc.

[0.4, -0.11, 0.55, 0.3 . . . 0.1, 0.02]

Others...

With increased adoption of pre-trained language models, “just the text” is the most common.



Tracking Multi-Turn…

The most important feature in modeling history is the **positional relationship between turns** to capture common patterns of conversational discourse.

Last-K context - a simple, but effective heuristic

Append the previous ***K*-context** (ctx) turns

- Previous user utterances (queries)
- Previous system utterances (responses)

Best *K* is typically 2-3

Effective across diverse models and tasks

- ConvQA models
- Dialog State Tracking
  (BERT w/3-ctx) – [Mehri et al., 2020]

	# contexts	CoQA	QuAC
BERT w/ 0-ctx	0	72.8	55.0
BERT w/ 1-ctx	1	79.2	63.4
BERT w/ 2-ctx	2	79.6	**65.4**
BERT w/ 3-ctx	3	79.6	65.3
BERT w/ 4-ctx	4	79.4	64.8
BERT w/ 5-ctx	5	**79.7**	64.5
BERT w/ 6-ctx	6	79.5	64.9
BERT w/ 7-ctx	7	**79.7**	64.4

Simple but Effective Method to Incorporate Multi-turn Context with BERT for Conversational Machine Comprehension [Ohsugi et al., 2019]

Context Dependence in ConvQA

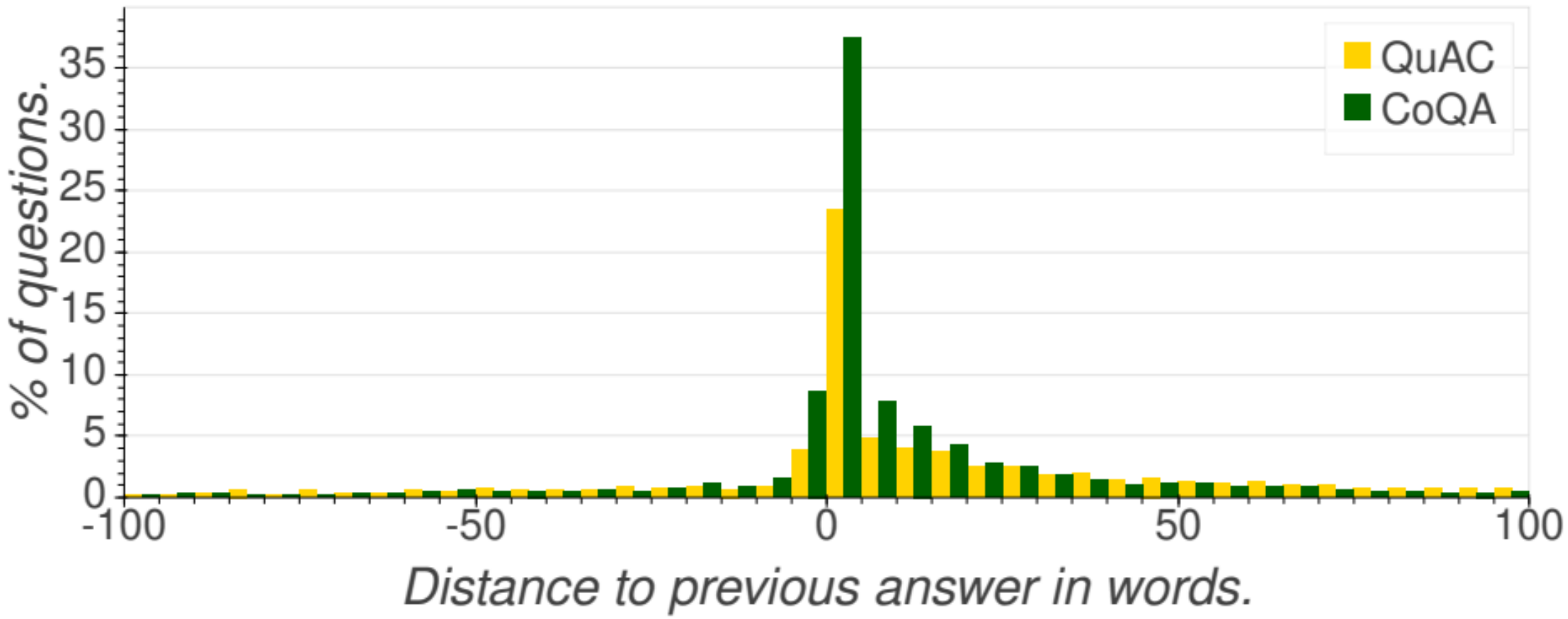


Figure from [Chiang et al. 2020]

Position Robustness Attack

Simply repeat the answer to make the distance between answers longer.

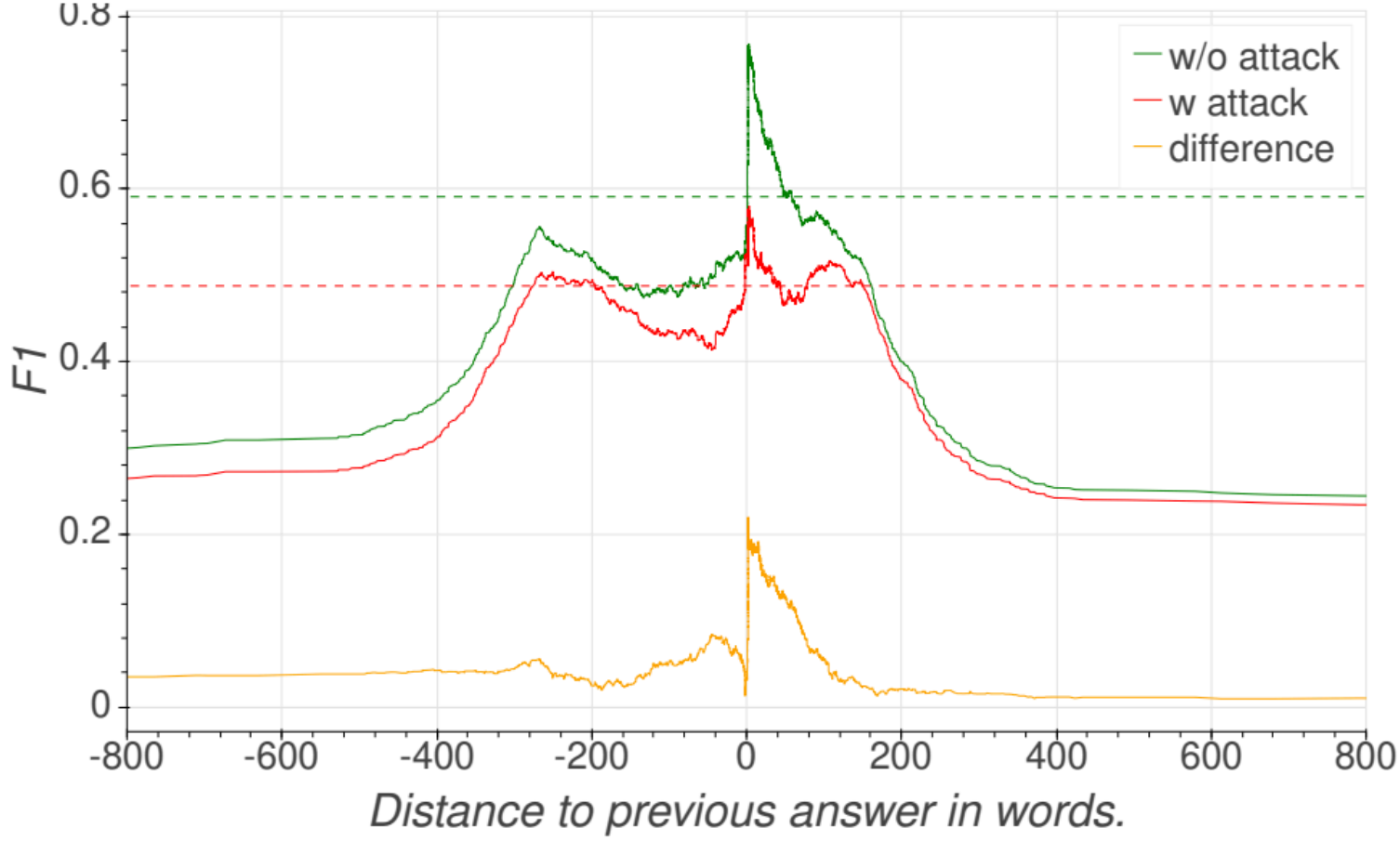


Figure from [Chiang et al. 2020]

Context Modeling Summary

- **Most context modeling approaches so far do not beat simple rules.**

Possible reasons

- Most datasets have 2-15 turns
- Single session on one topic
- Most turns have only ‘local’ dependence on context

Beyond *k-ctx* → advanced models of history

- Open research area
- May require new conversational collections

Conversational Language Understanding Tasks

Conversational Query Expansion (CQE)

Conversational Query Rewriting (CQR)

Conversational Entity Detection and Linking (CEDL)

Context Salience (word, turn)

Discourse & Sub-topic Classification

Context Ranking & Summarization

Query Expansion

Task

$$Q \rightarrow Q_{exp}$$

Select and weight important words from the conversation history and/or PRF.

- **Rules**
  - First turn, previous turn [Clarke, 2019]
  - Historical Query Expansion (HExp) – [Yang et al., 2019]
- **Supervised**
  - Conversational Term Selection (CVT) [Kumar et al., 2020]
  - QuReTeC [Voskarides et al., 2020]

Algorithm 1: Historical Query Expansion

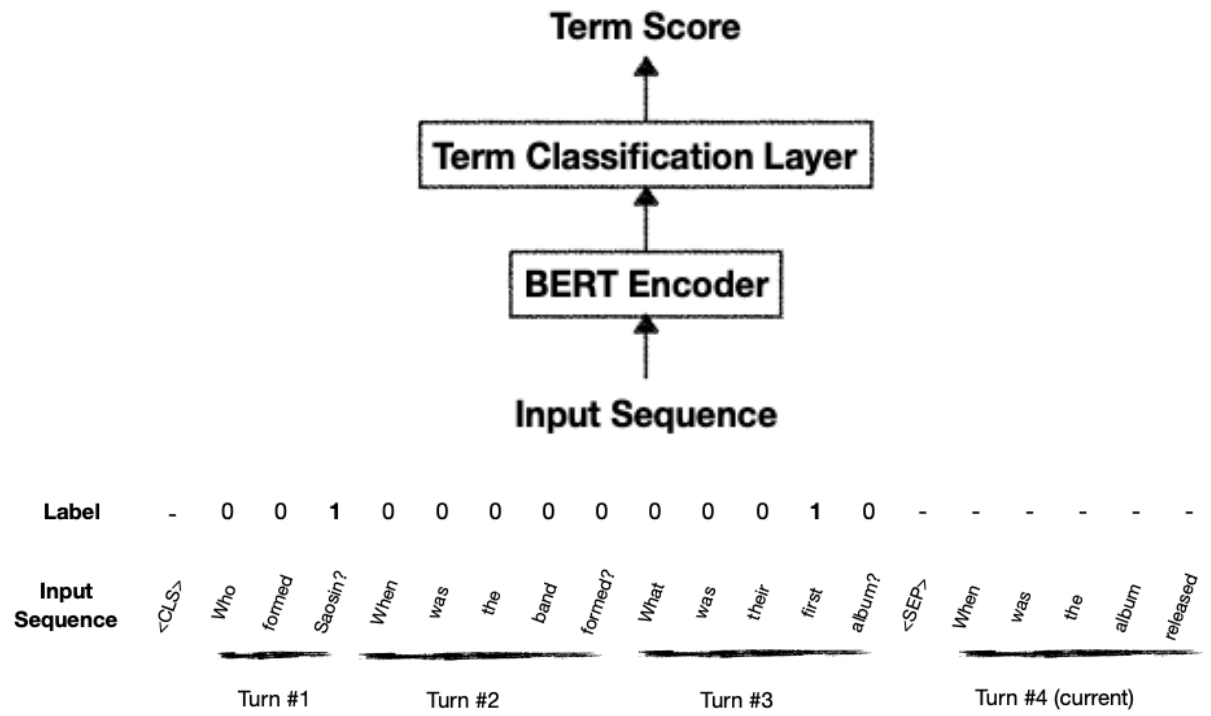
Input: $u_i, u_{<i}, C$
Output: $\bar{u}_i$

$\bar{u}_i \leftarrow ()$; $W_{\text{topic}} \leftarrow \{\}$; $W_{\text{sub}} \leftarrow \{\}$
for $j = 1$ *to* i **do**
  for $k = 1$ *to* $n(u_j)$ **do**
    $\mathcal{R}_j^k = \text{KE}\left(t_j^k, C\right)$
    if $\mathcal{R}_j^k > \mathcal{R}_{\text{topic}}$ **then**
      $W_{\text{topic}}.\text{insert}\left(t_j^k\right)$
    if $(\mathcal{R}_j^k > \mathcal{R}_{\text{sub}})$ **and** $(j \geq i - \mathcal{M})$ **then**
      $W_{\text{sub}}.\text{insert}\left(t_j^k\right)$
if $i > 1$ **then**
  $\mathcal{A}_i = \text{QPP}(u_i, C)$
  $\bar{u}_i.\text{insert}(t)$ for all $t \in W_{\text{topic}}$
  if $\mathcal{A}_i < \eta$ **then**
    $\bar{u}_i.\text{insert}(t)$ for all $t \in W_{\text{sub}}$
$\bar{u}_i.\text{append}(u_i)$
return $\bar{u}_i$

Figure from [Yang et al., 2019]

Supervised Query Expansion

- Query Resolution by Term Classification (QuReTeC)
- BERT binary term classification from the history
- Labels from distant supervision from rel. passages



(b) Example input sequence and gold standard term labels (1: relevant, 0: non-relevant) for QuReTeC.

Table 10: Qualitative analysis for initial retrieval (extrinsic) when using QuReTeC or RM3 (cur+first) for query resolution. The example is sampled from the TREC CAsT dataset.

Q1: What is a real-time database?
Q2: How does it differ from traditional ones?
Q3: What are the advantages of real-time processing?
Q4: What are examples of important ones?
Q5: What are important applications?
Q6: What are important cloud options?
Q7: Tell me about the Firebase DB?
Q8 (current): How is it used in mobile apps?

Predicted terms – QuReTeC: {"database", "firebase", "db" }

Top-ranked passage – QuReTeC

Firebase is a mobile and web application platform …Firebase's initial product was a realtime database, …Over time, it has expanded its product line to become a full suite for app development

Predicted terms – RM3 (cur+first): {"real", "time", "database"}

Top-ranked passage – RM3 (cur+first)

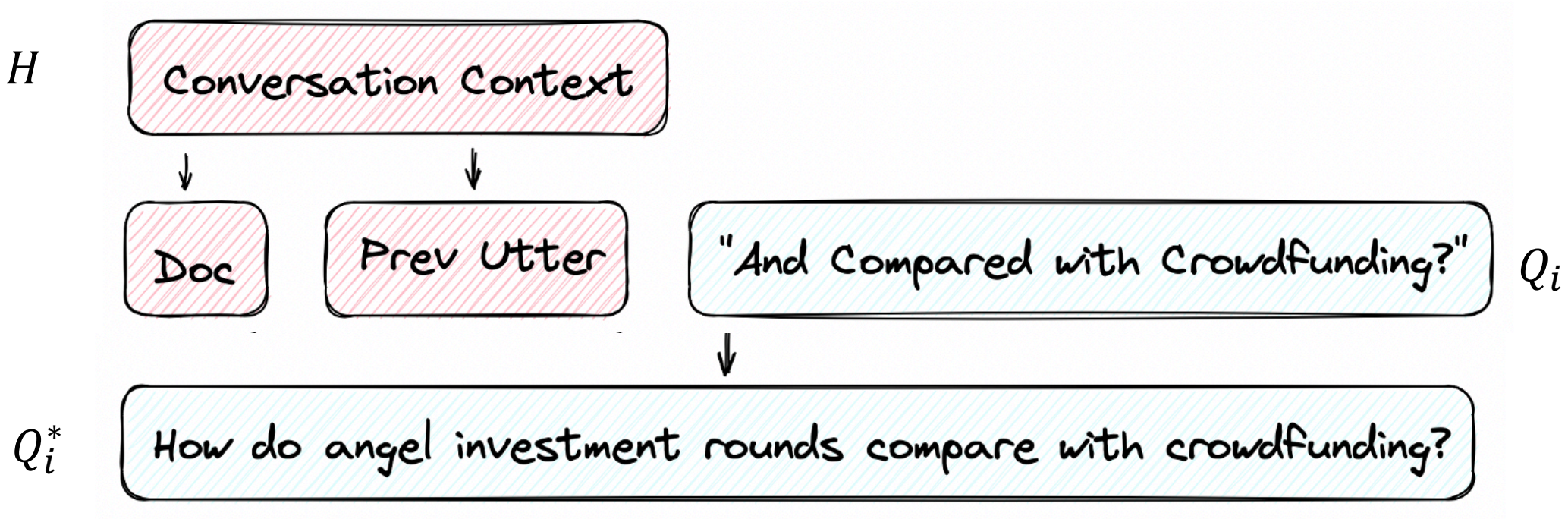
There are two options in Jedox to access the central OLAP database and software functionality on mobile devices: Users can access reports through the touch-optimized Jedox Web Server …on their smart phones and tablets.

Figures from [Voskarides, et. al, 2020]

Conversational Query Rewriting (CQR)

- Generate contextualized query from conversation context
- Sequence-to-sequence approach = $P(\boldsymbol{y}|\boldsymbol{x};\ \theta) = \prod_{j=1}^{J} P\left(y_j \middle| y_{<j}, x; \theta\right)$

$$Q_i^* = CQR(Q_i, H;\ \theta)$$



CAsT Y1: NLP-based rewriting baseline

- Run AllenNLP or other NLP toolkit to identify entities and mentions.
- **Rewriting**: Replace ‘coreferent’ mentions with ‘canonical name’

How much does 0 a used Lamborghini cost ? How does 0 it compare to a Ferrari ? Interesting . What about for a pimped – out food truck ? What licenses and permits are needed ? What is a typical day like ? How can I run 1 it successfully ? What are some good examples to learn from ? Besides inventive flavors , what made 1 it successful ?

CAsT Y1 Coreference phenomena

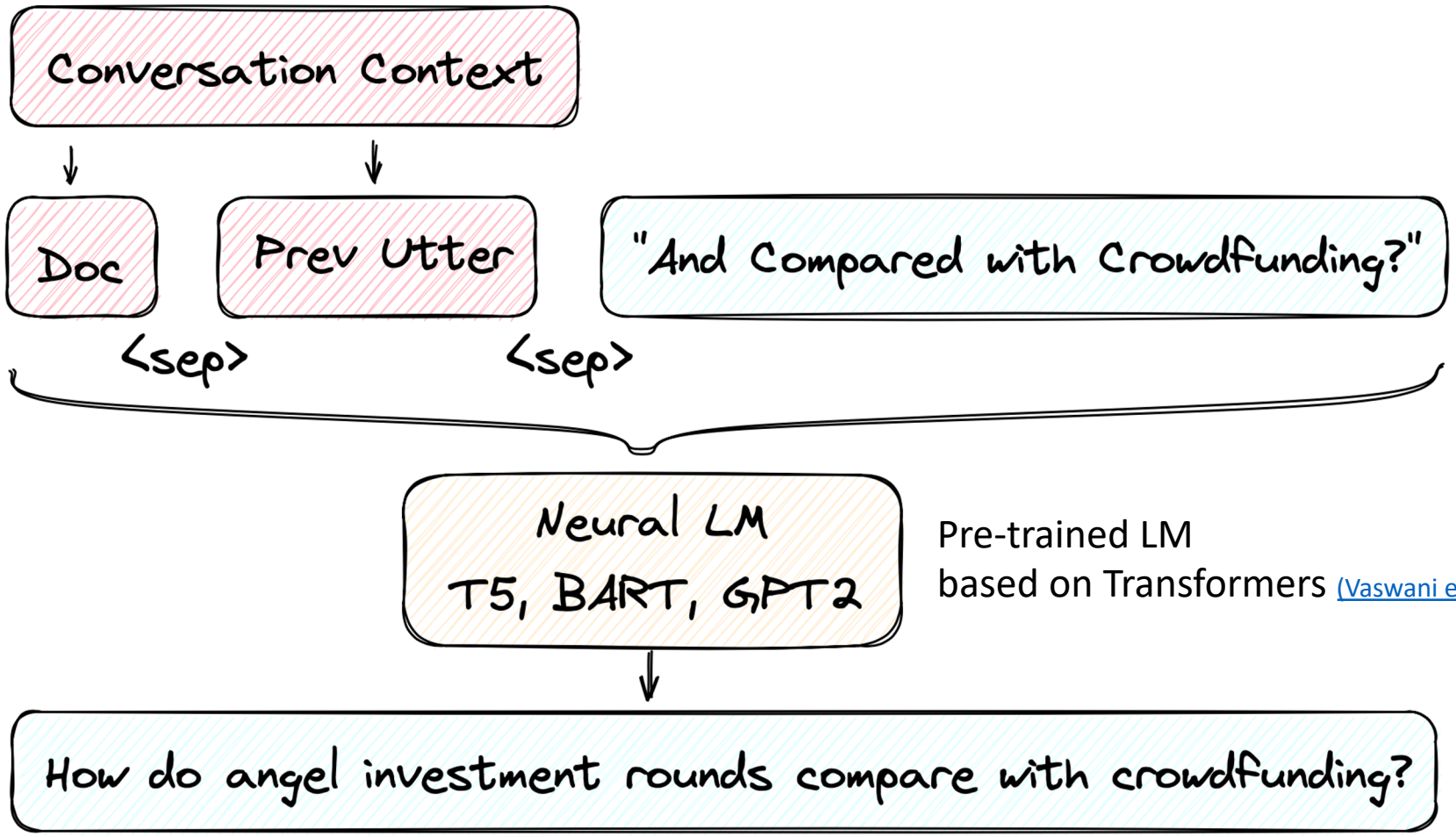
Examples

Type	Utterance	Mention
Pronominal	How do **they** celebrate Three Kings Day?	they -> Spanish people
Zero	What cakes are traditional?	Null -> Spanish, Three Kings Day
Groups	Which **team** came first?	which team -> Avengers, Justice League
Abbreviations	What are the main types of **VMs**?	VMs -> Virtual Machines

Statistics

Dataset	Pronominal	Zero	Groups	Abbreviations
TRAIN	102	**82**	6	29
EVALUATION	128	**111**	4	15

Neural Query Rewriting



Pre-trained LM
based on Transformers (Vaswani et al. 2017)

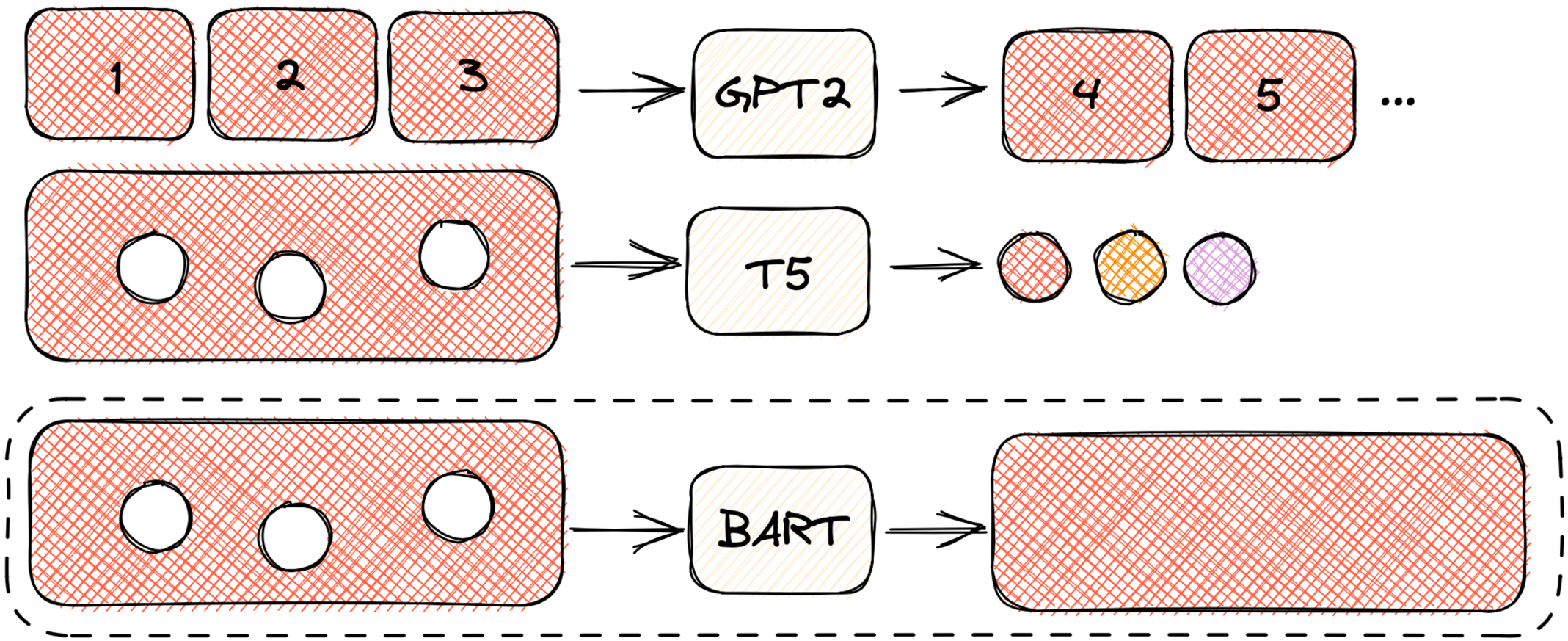
CQR Datasets

- **CAsT (2019-present)**
  - Few-Shot Generative Conversational Query Rewriting [Yu et al., 2020]
  - A few hundred turns

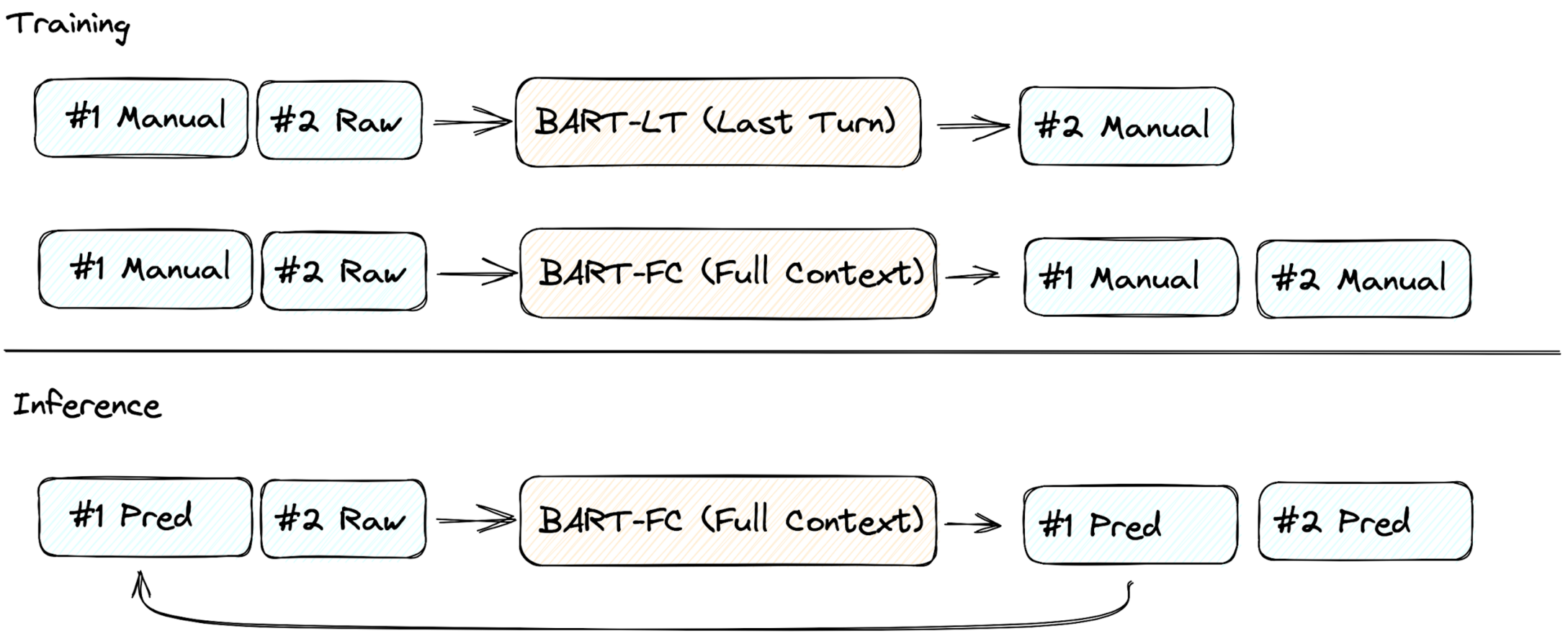
- **CANARD – (2019)**
  - Based on QuAC - 40,527 turns
  - Conversational Question Reformulation via Sequence-to-Sequence Architectures and Pretrained Language Models [Lin et al., 2020]

- **QReCC (2021)**
  - 14K conversations with 81K question-answer pairs

Neural LM Pre-Training

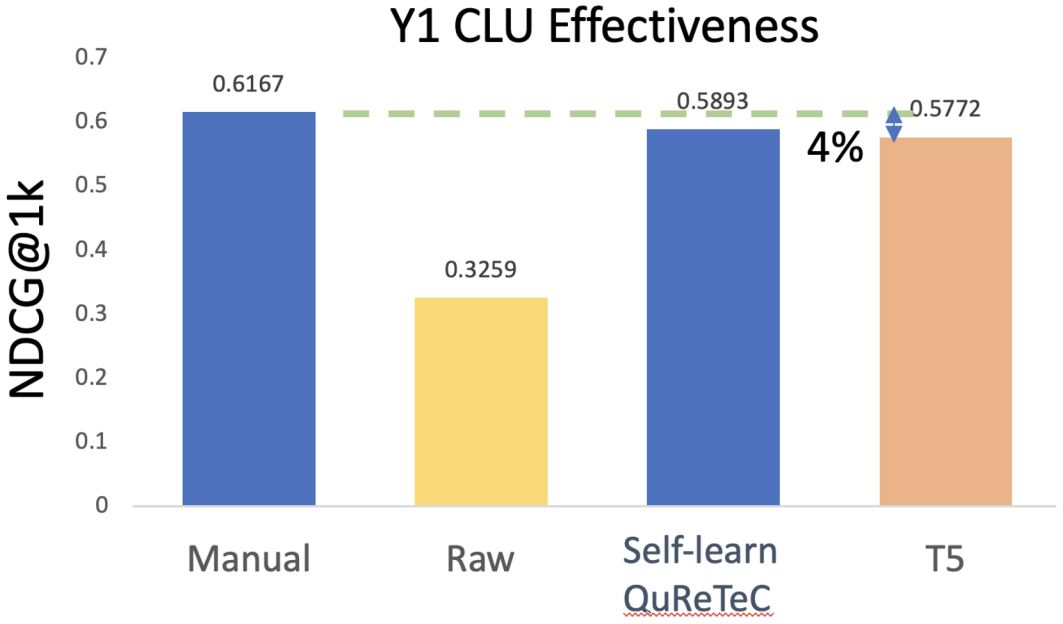


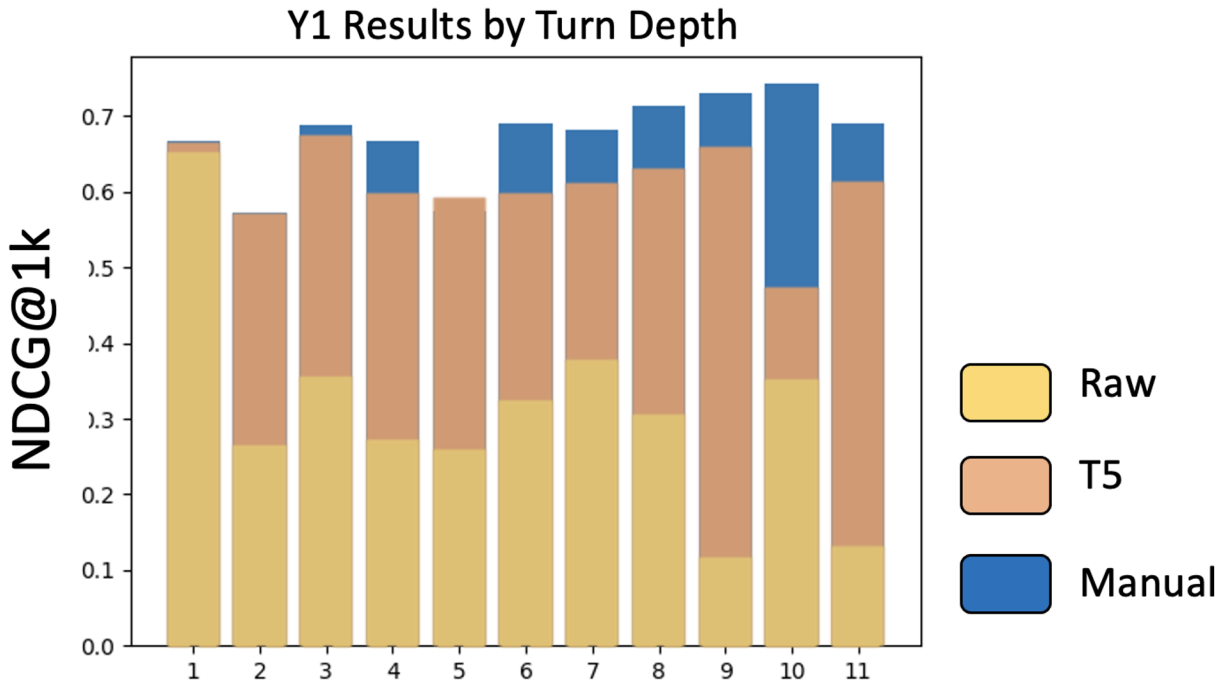
BART FC: Aligning Pre-Training with Fine-Tuning



[Gemmell et al. 2020]

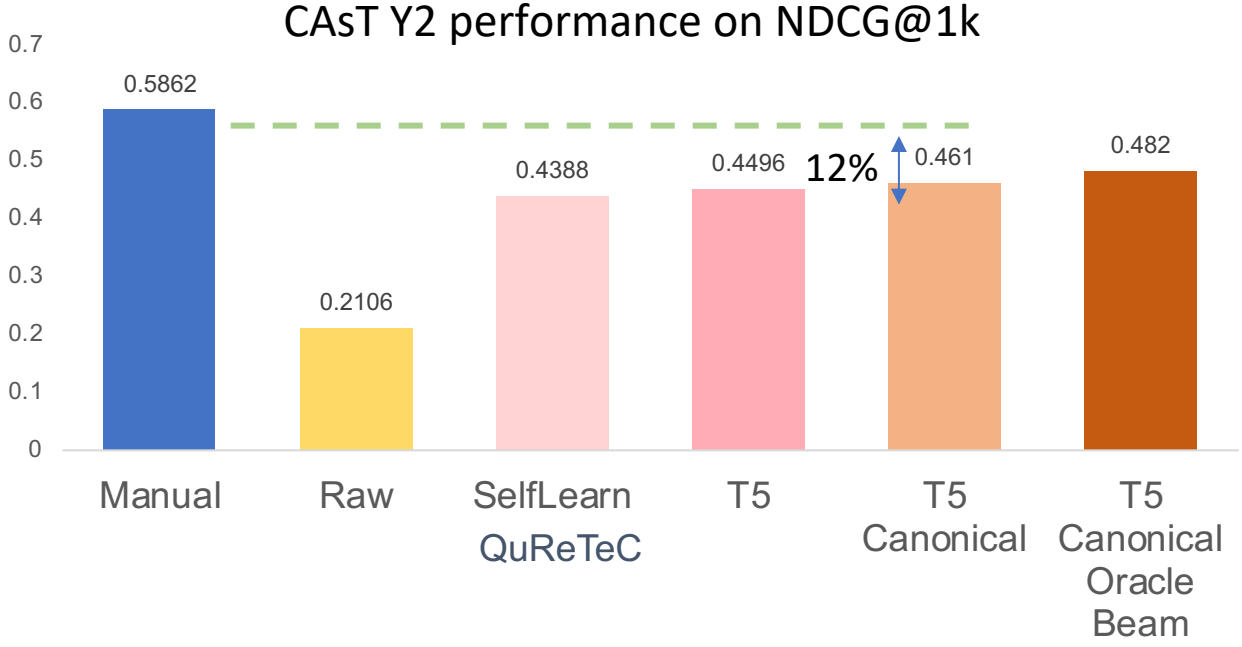
CAsT Y1 Results: Rewriters Do Well

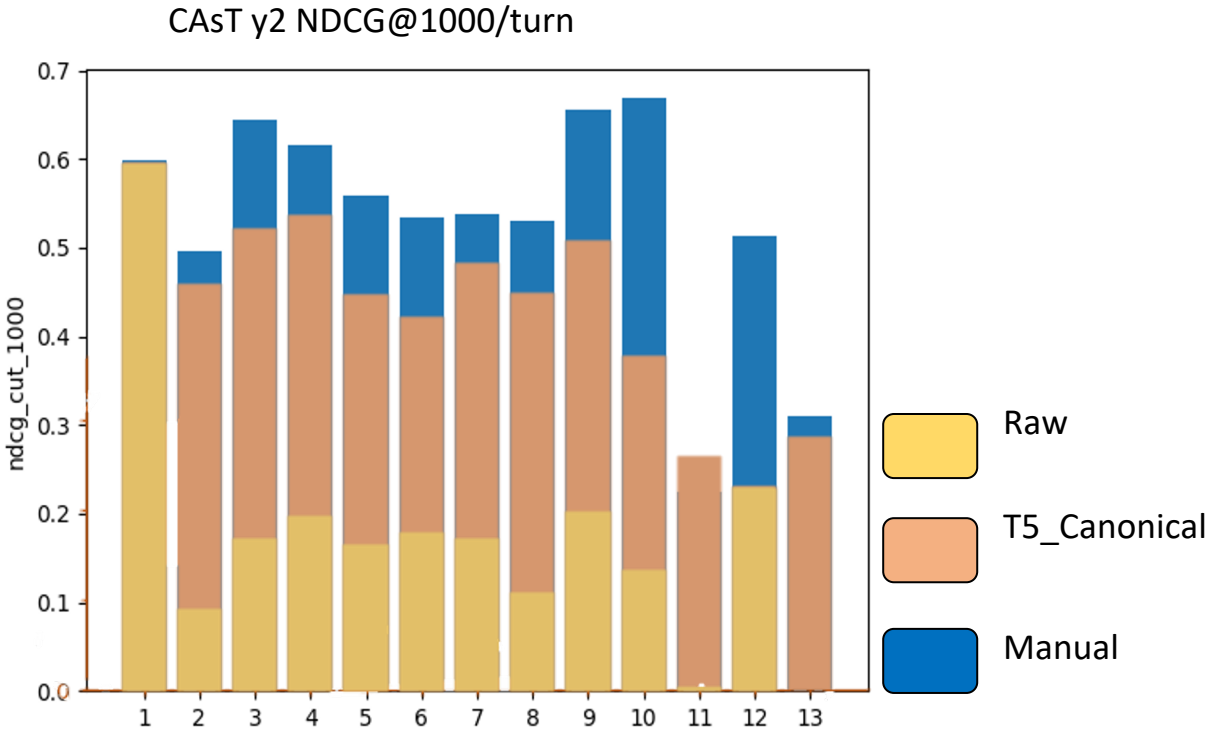




No dependence on results, only previous user queries.

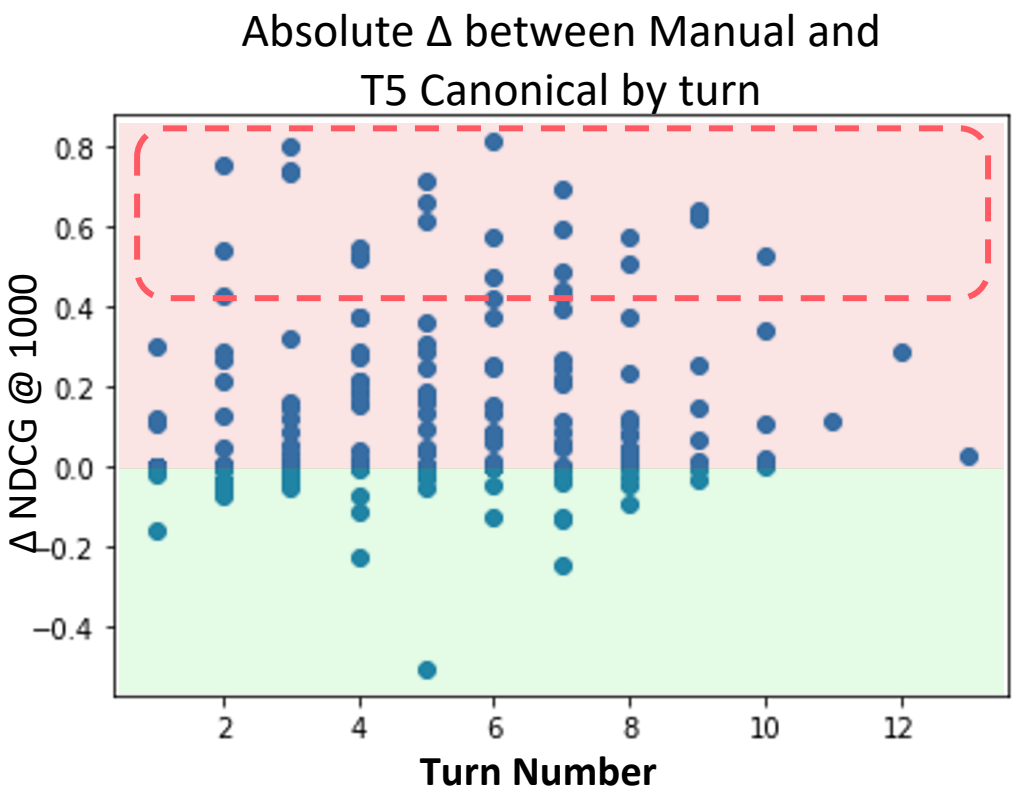
CAsT Y2 Results: Rewriters Struggle

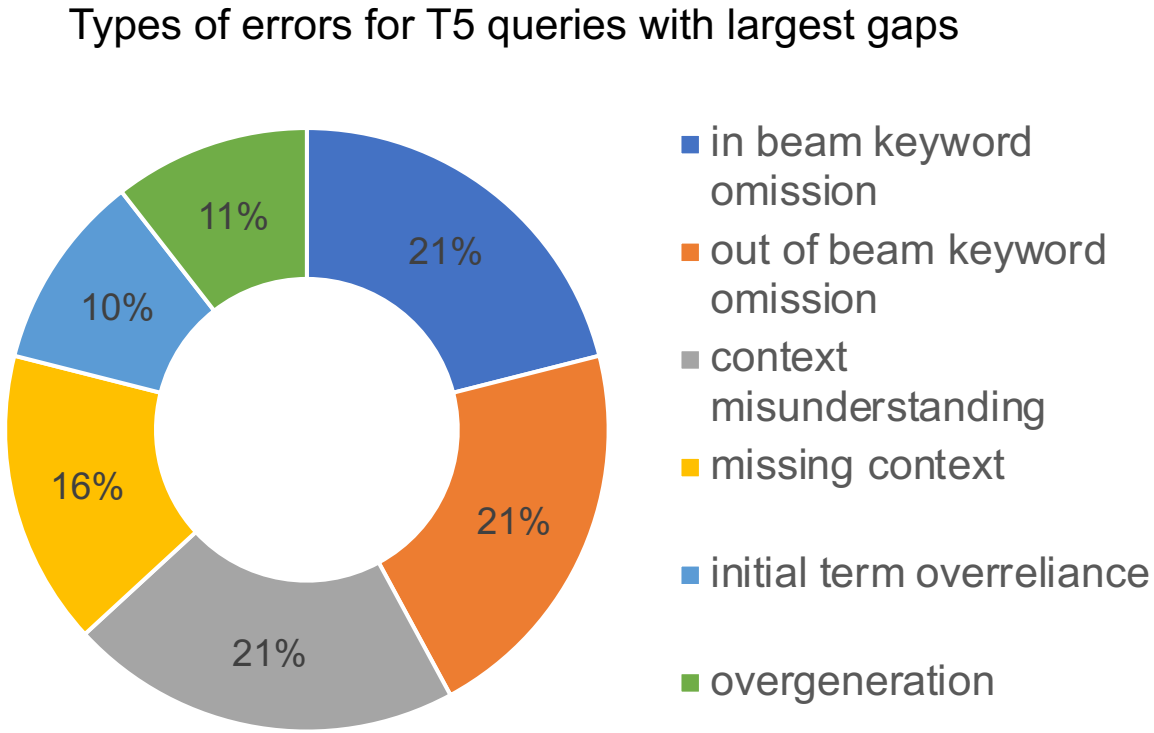




Adds long response result dependence

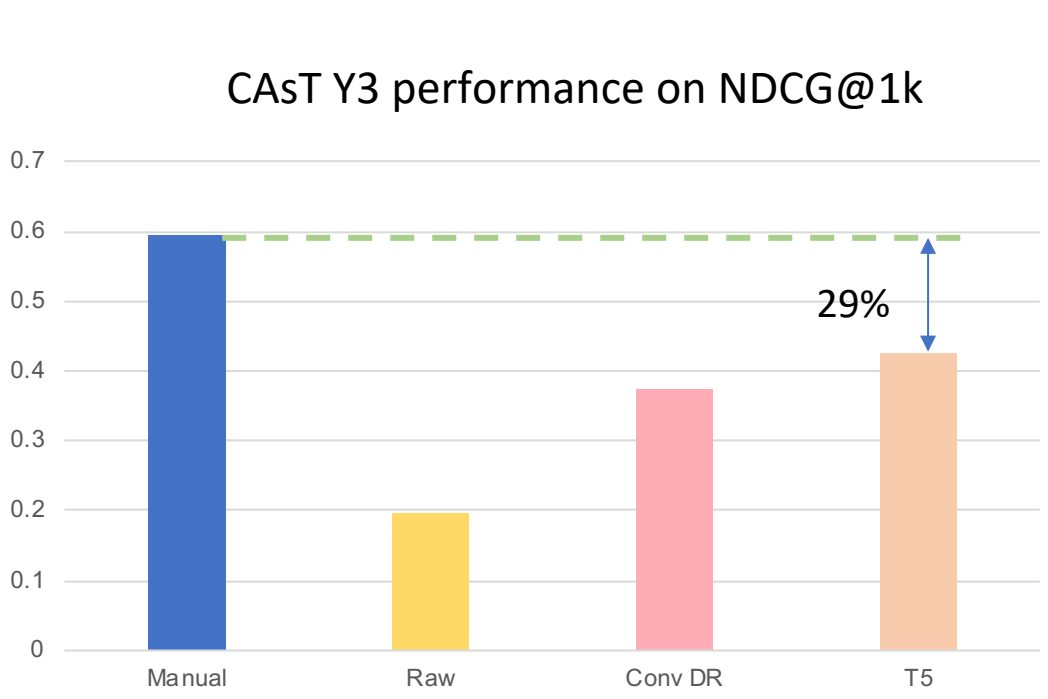
CAsT Analysis: Year 2 Rewriters Break Down

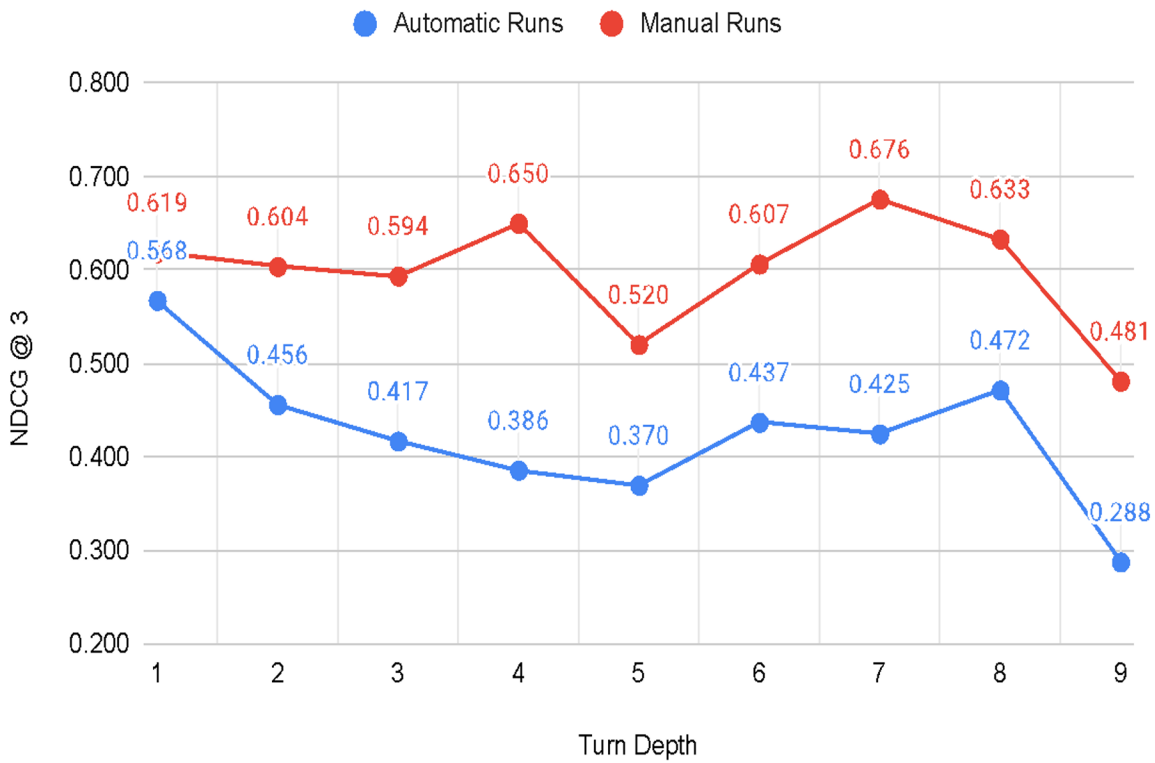




Rewrite failure can lead to catastrophic first phase retrieval failure.

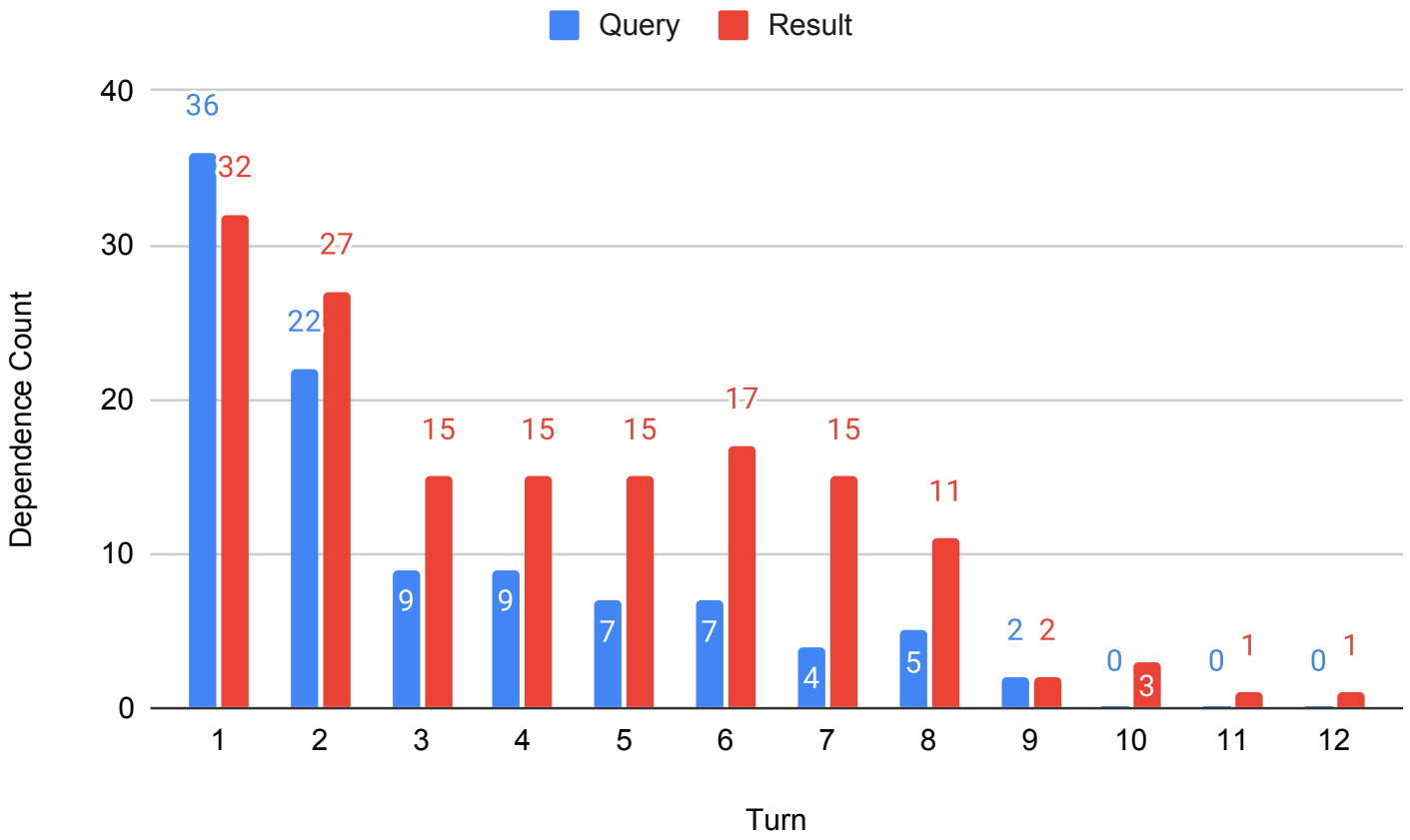
CAsT Y3 Results: Rewriters Struggle More





More complex dependence of varying structure

Context Dependence in CAsT Y3



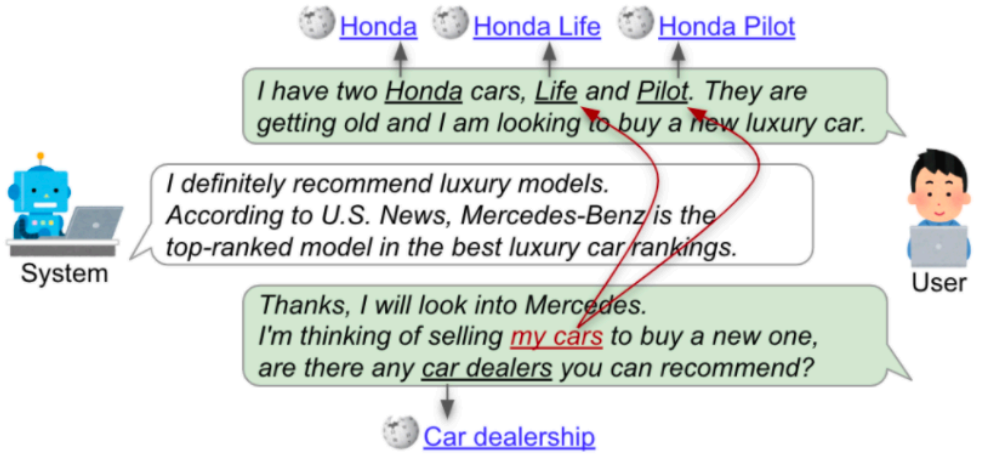
‘Non-trivial’ dependence

→Dependence with > 1 turn distance [Dalton et al., 2021]

Conversational Entity Detection and Linking (CEDL)

Goal

Recognize and link entities (named entities and concepts) to a knowledge base



Example from [Joko et al. 2022]

A few recent findings:

- Mentions of personal entities are mainly in social chat conversations
- Concepts beyond named entities are important for conversational intent modeling
- Traditional entity linking approaches fall short, but new datasets are making progress

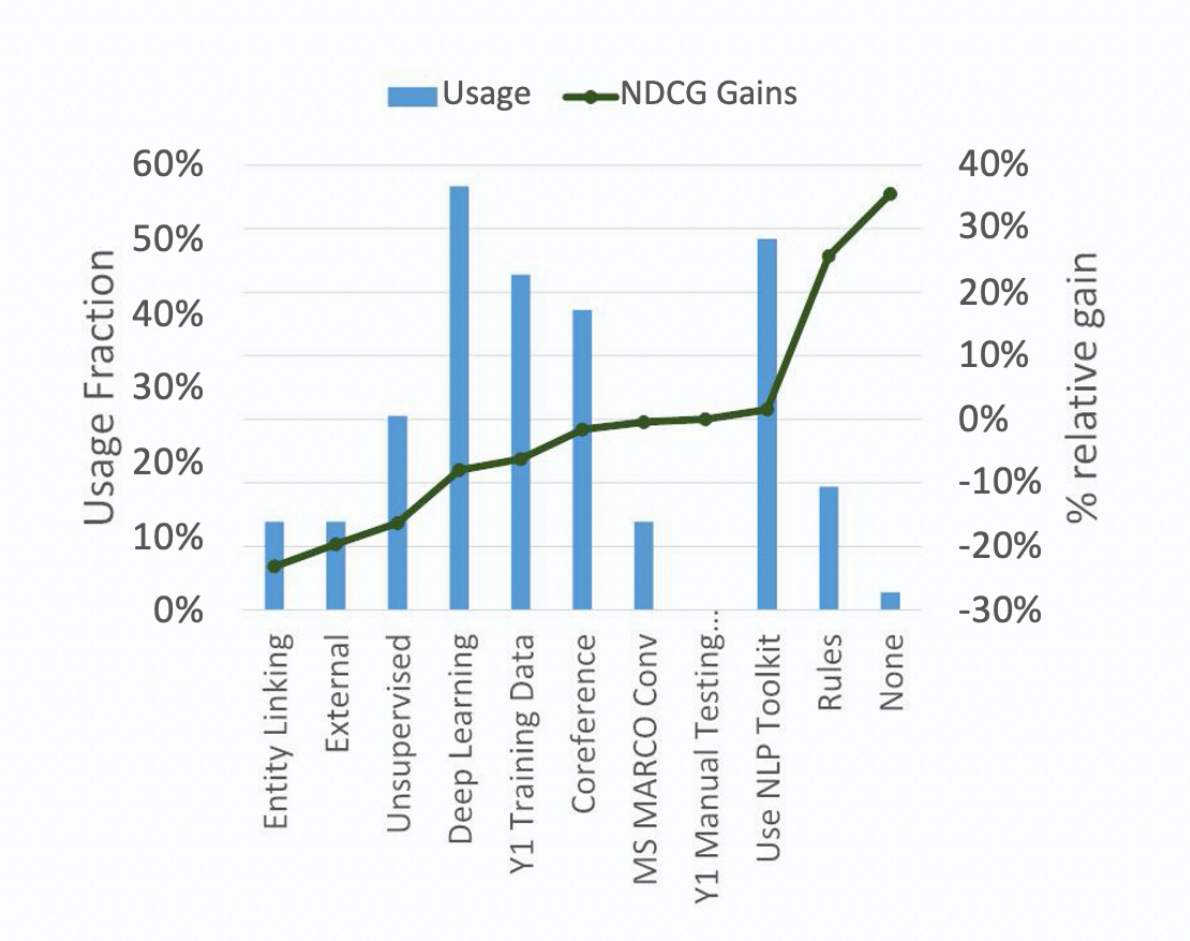
	ConEL-2 (Val)		ConEL-2 (Test)		ConEL	
	F_{MD}	F_{EL}	F_{MD}	F_{EL}	F_{MD}	F_{EL}
GENRE [3]	.290	.252	.320	.299	.350	.211
REL [29]	.304	.244	.279	.231	.462	.245
TagMe [8]	.559	.478	.611	.504	.510	.375
WAT [22]	.616	.539	.613	.519	.416	.336
REL*($BERT_{conv}$)	**.748**	**.652**	.716	.587	.551	.424
REL*($BERT_{WP\text{-}conv}$)	.697	.624	.708	.592	.508	.399
REL*($BERT_{NER\text{-}conv}$)	.723	.635	.693	.576	.548	.407
CREL	.742	.651	**.729**	**.597**	**.559**	**.429**

[Joko et al. 2022]

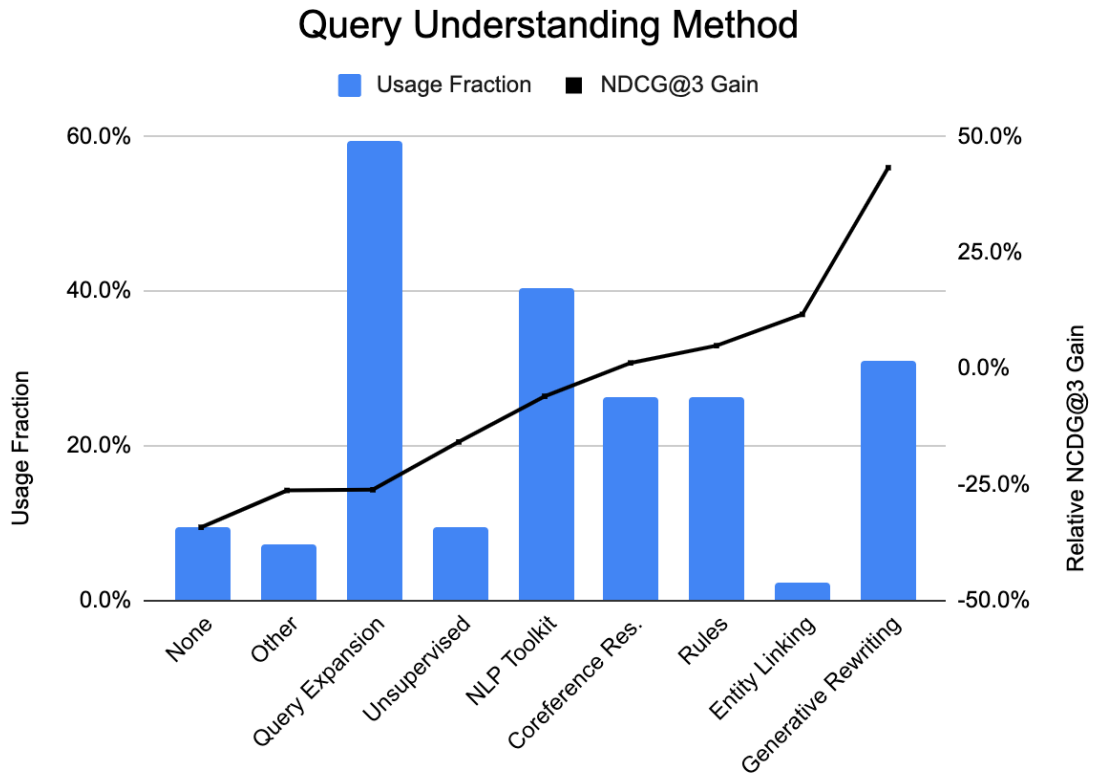
https://github.com/informagi/conversational-entity-linking-2022

Putting the pieces together…

- Evolution of Conversational Language Understanding methods in CAsT



2019



2020

Putting the pieces together

Effectiveness requires combining conversational tasks

- Conversational Query Rewriting
- Conversational Query Expansion ...

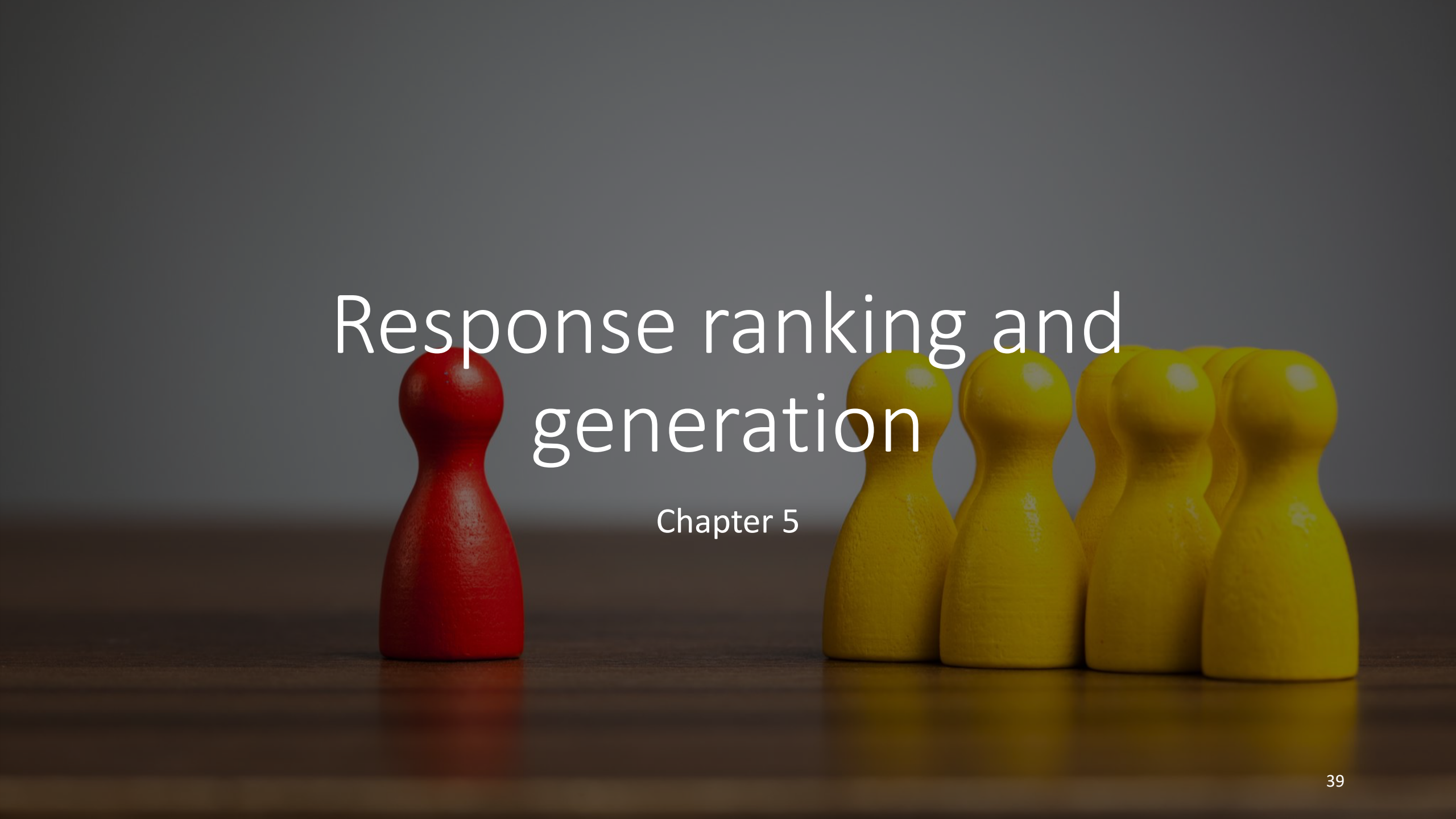
CLU may be performed in multiple phases in combination with with multi-stage pipelines

[Lin et al., 2021]

Need new methods for generating effective conversational representations beyond current methods

Response ranking and generation

Chapter 5



Ranking and Generation Tasks

Short answers

ConvQA
KG-ConvQA
OR-ConvQA

Long answers

Conv PR
Conv DR

Semi-structured data

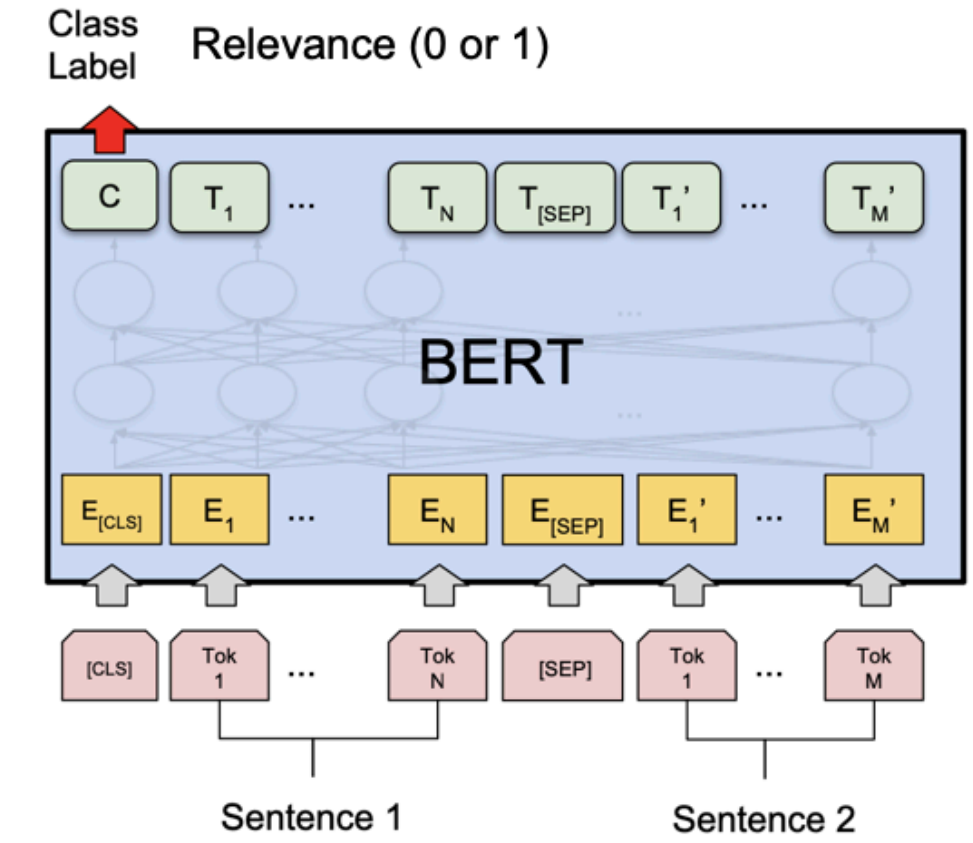
Conv TR

Recommendation

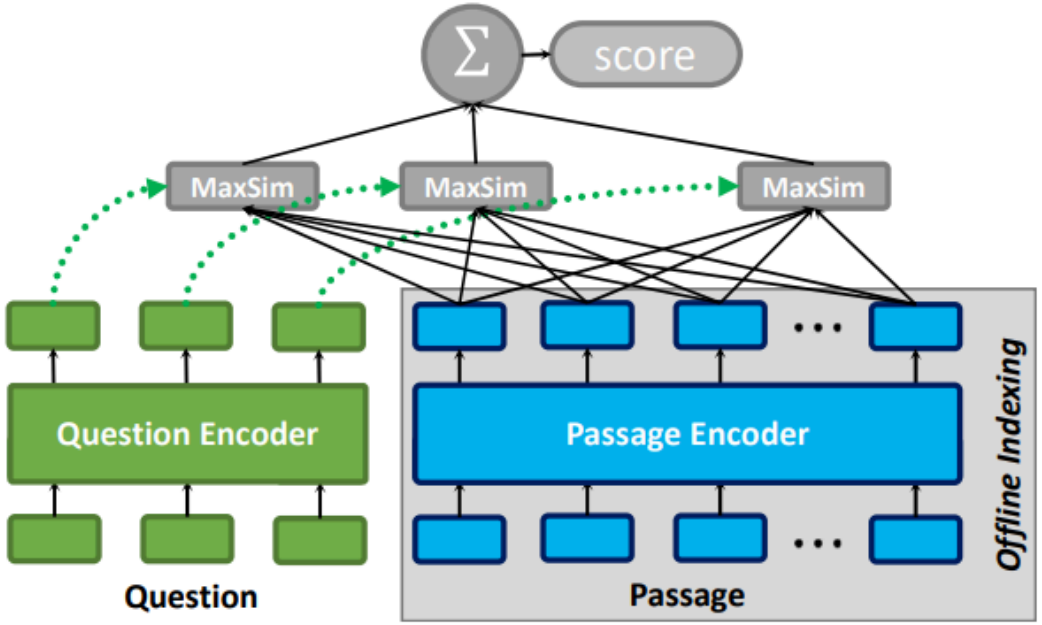
ConvRec

Widespread use of Neural LMs for ConvQA and Retrieval

Pre-trained BERT large uncased model fine-tuned on MS MARCO from [Nogueira and Cho, 2019]



ColBERT



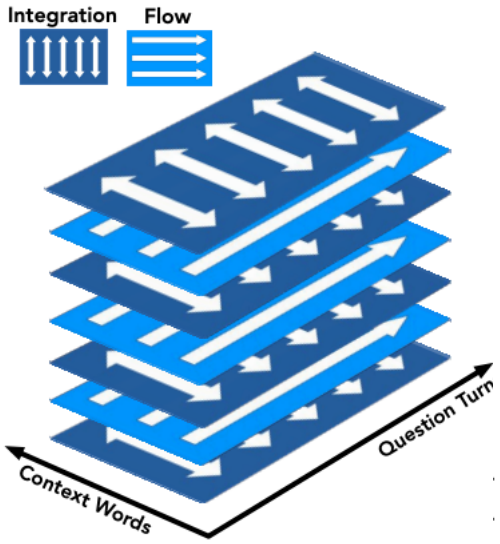
[Khattab et al. 2020]

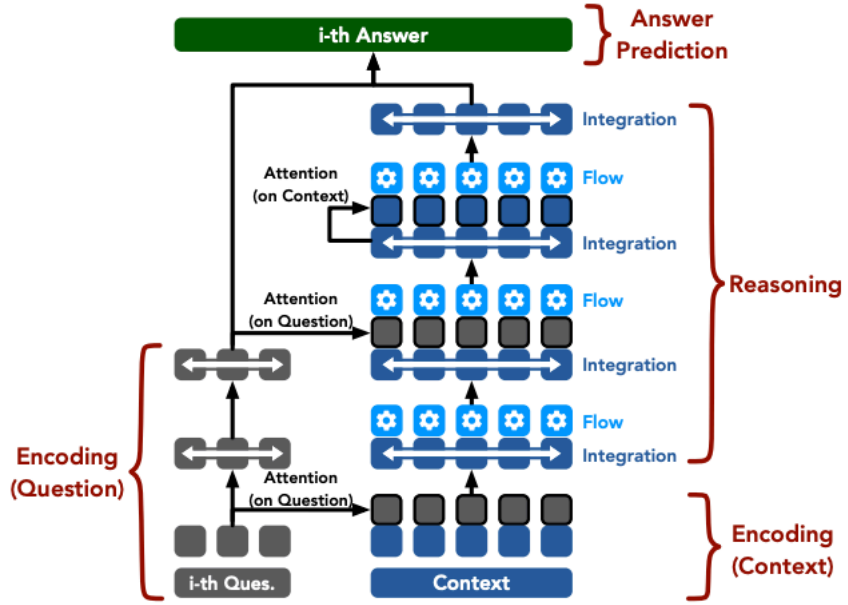
ConvQA Evolution

- TREC 2004 – Contextual QA w/coref in a pipeline
- ConvQA – Machine Comprehension QA models w/context
  - DrQA, BiDAF++
- OR-ConvQA – Retrieve and Read Models
  - BERT-based – BERTserini (retriever) + BERT QA (reader)
  - Fusion-in-Decoder – Dense Passage Retriever + T5 reader [Izacard and Grave, 2021]

Flow-based models for ConvQA

- Add layers for reasoning about turn-level evidence
- **Information Flow** layers
  - **Integration** - For each turn, create a integrated contextualized vector for every word
  - **FLOW** layer – Integrate token-level contexts across turns
- FlowDelta - Explicitly model information gain in conversation
- GraphFlow – Model FLOW using RGNNs





[Huang et al., 2019]

Q1: Who went to the farm? -> Q2: Why?

Billy went to the farm to buy some beef for his brother 's birthday . When he arrived there he saw that all six of the cows were sad and had brown spots . The cows were all eating their breakfast in a big grassy meadow . He thought that the spots looked very strange so he went closer to the cows to get a better look …

Q2: Why? -> Q3: For what?

Billy went to the farm to buy some beef for his brother 's birthday . When he arrived there … After Billy got a good look at the cows he went to the farmer to buy some beef . The farmer gave him four pounds of beef for ten dollars . Billy thought that …

[Chen et al., 2020]

ConvQA with Transformers - BERT

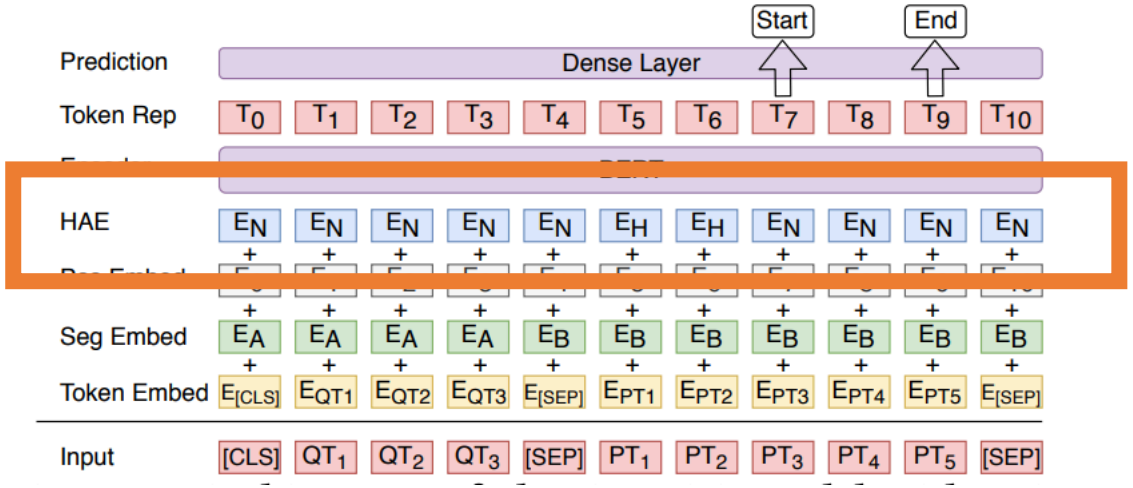
AllHistory
- Append turns with separators

```
[q1 [SEP] a1 [SEP] q2 [SEP] a2 [SEP]...
[SEP] qn-1 [SEP] an-1 [SEP] qn.]
```

[Adlakha et al. 2021]

History Answer Embedding

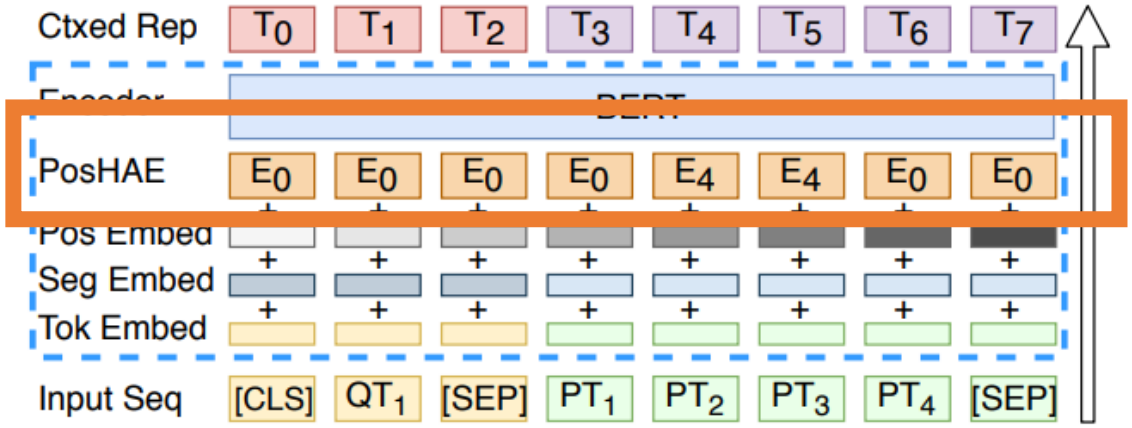
- Distinguishes between user/response in context using a binary embedding.
- 63.9% F1 on QuAC



[Qu et al., 2019a]

Positional History Answer Embedding

- A shared **relative** position indicator.
- 64.7% on QuAC

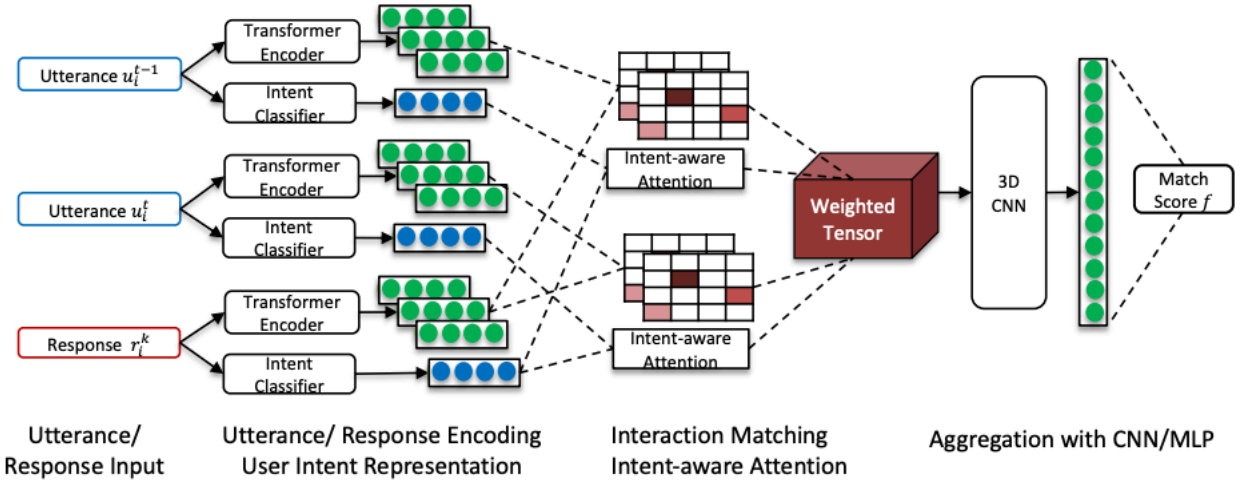


[Qu et al., 2019b]

Beyond Answers: Discourse-Aware Ranking

- Defines a custom taxonomy of user information intents
- Given predicted intent, generates intent-aware matching features
- A stacked 3D CNN layer to combine weighted evidence
- A step towards a model with rich support for more complex types of interaction

Code	Label	Description
OQ	Original Question	The first question that initiates a QA dialog
RQ	Repeat Question	Questions repeating a previous question
CQ	Clarifying Question	Users or agents ask for clarification
FD	Further Details	Users or agents provide more details
FQ	Follow Up Question	Follow-up questions about relevant issues
IR	Information Request	Agents ask for information from users
PA	Potential Answer	A potential solution to solve the question
PF	Positive Feedback	Positive feedback for working solutions
NF	Negative Feedback	Negative feedback for useless solutions
GG	Greetings/Gratitude	Greet each other or express gratitude
JK	Junk	No useful information in the utterance
O	Others	Utterances that cannot be categorized

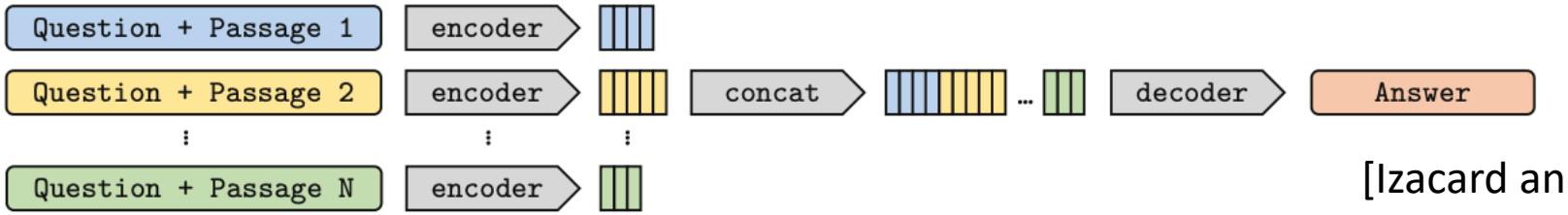


[Yang et. al, 2020] - IART: Intent-aware Response Ranking

Generative OR-ConvQA with Transformers

Fusion-in-Decoder

Retrieve (DPR) + Generative QA model (T5)



[Izacard and Grave, 2021]

Model	Question Rep	Dev		Test	
		EM	F1	EM	F1
Human		**40.2**	**70.1**	**40.3**	**70.0**
GPT-3		12.4	33.4	10.4	31.8
BM25 + DPR Reader	ORIGINAL	7.1	12.8	7.2	13.0
	ALLHISTORY	13.6	25.0	13.8	25.2
	REWRITES	15.4	32.5	15.7	31.7
BM25 + FiD	ORIGINAL	10.1	21.8	10.5	22.6
	ALLHISTORY	24.1	37.2	23.4	36.1
	REWRITES	24.0	41.6	24.9	41.4
DPR Retriever + DPR Reader	ORIGINAL	4.9	14.9	4.3	14.9
	ALLHISTORY	21.0	43.4	19.4	41.1
	REWRITES	17.2	36.4	16.5	35.2
DPR Retriever + FiD	ORIGINAL	7.9	21.6	7.8	21.4
	ALLHISTORY	**33.0**	**55.3**	**33.4**	**55.8**
	REWRITES	23.5	44.2	24.0	44.7

Table 5: Overall performance of all model variants on TOPIOCQA development and test set

Results from [Adlakha et al., 2021]

- Rewriting is particularly important in initial retrieval; less important for Reader model

KG-ConvQA – CONVEX to CONQUER

- Start from a seed entity and perform actions to traverse the KG

CONVEX [Christmann et al. 2019]

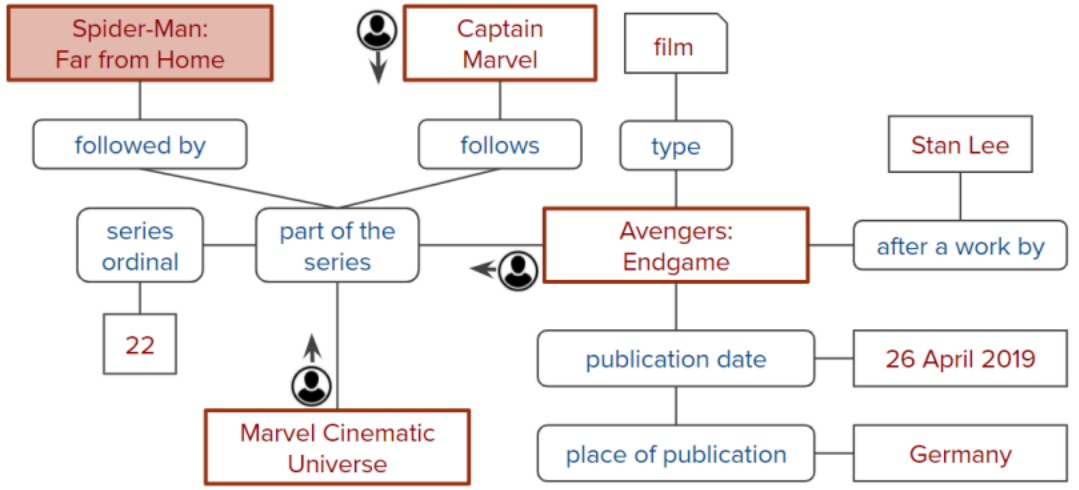
- Judiciously expand a graph within a k-hop neighborhood based on:
  1) Relevance to question
  2) Relevance to context
  3) KG priors

CONQUER [Kaiser et al. 2021]

- Select KG graph actions using RL
- Support for feedback/reformulation

Examples from [Kaiser et al. 2021]

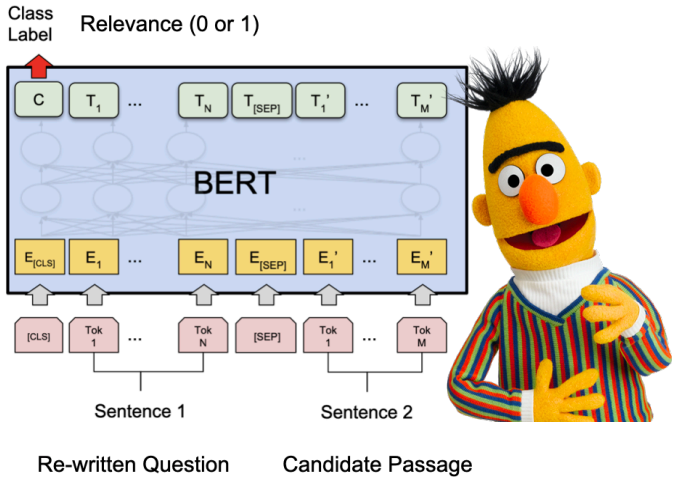
q_1: *When was Avengers: Endgame released in Germany?*
ans_1: *24 April 2019*
q_{21} : *What was the next from Marvel?* **(New intent)**
ans_{21}: *Stan Lee* **(Wrong answer)**
q_{22} : *What came next in the series?* **(Reformulation)**
ans_{22}: *Marvel Cinematic Universe* **(Wrong answer)**
q_{23}: *The following movie in the Marvel series?* **(Reformulation)**
ans_{23}: *Spider-Man: Far from Home* **(Correct answer)**
q_{31}: *Released on?* **(New intent)**



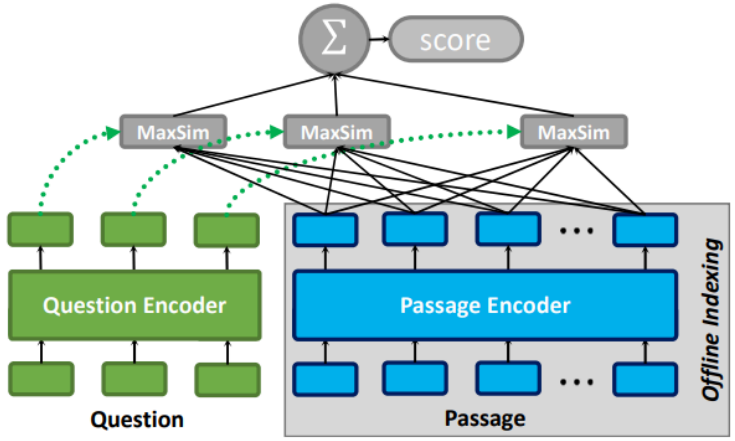
ConvPR and ConvDR

ConvPR with Rewritten queries

BERT for neural ranking

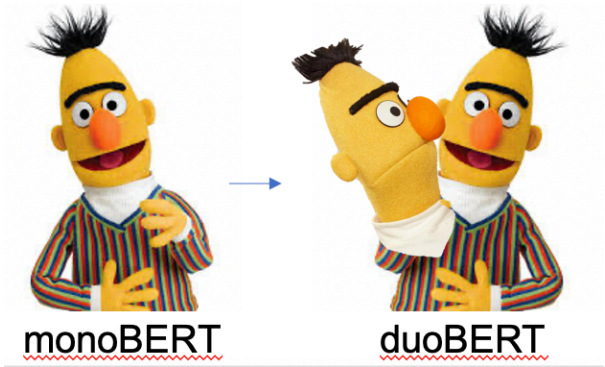


ColBERT



[Khattab et al. 2020]

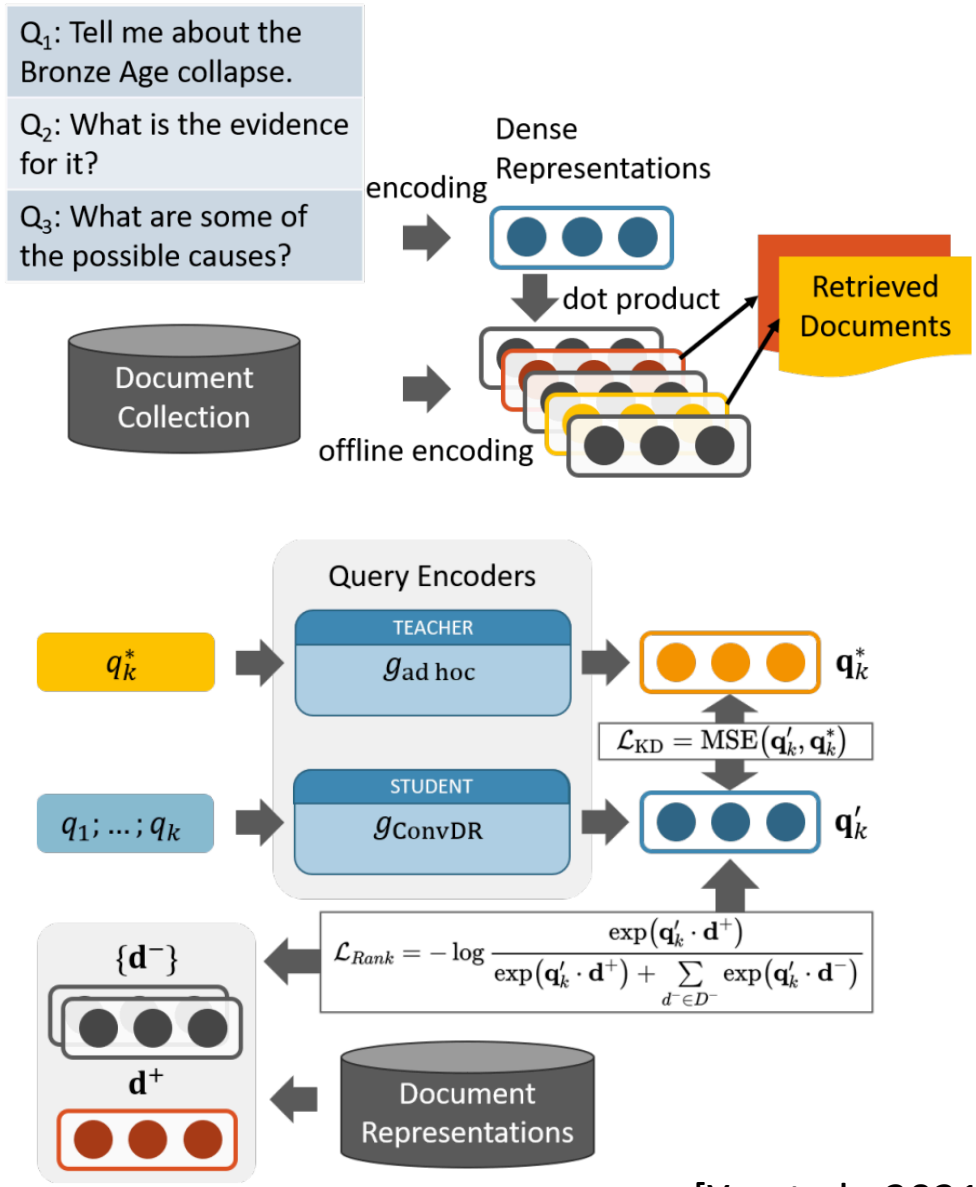
Multi-stage fusion with neural ranking



T5

Conversational Dense Retrieval

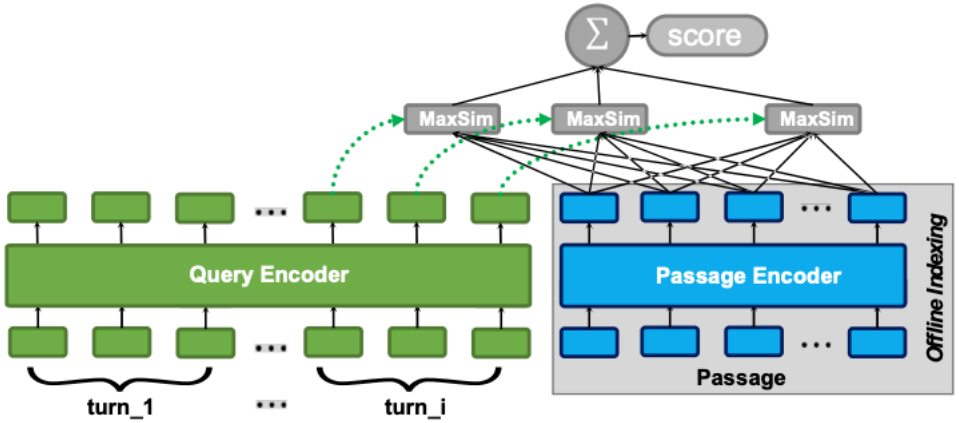
- End-to-end retrieval with dense vector representation to retrieve content
- Can be more effective than rewriting + dense retrieval
- Learned from oracle manual query representations



[Yu et al., 2021]

Zero-shot Conversational Contextualization

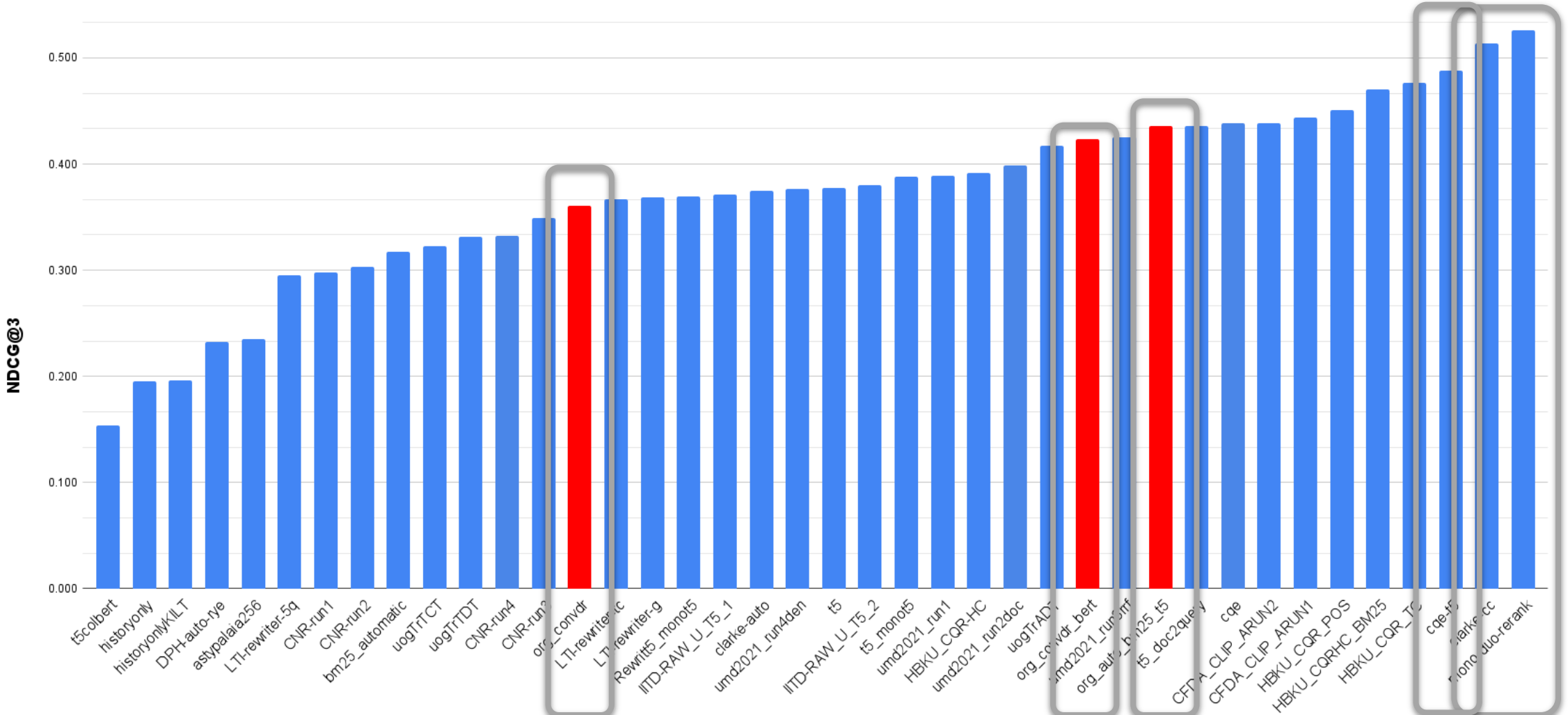
- $ZeCo^2$ -A variant of ColBERT for ConvPR
  - Contextualizes all embeddings within the conversation
  - Matches only the contextualized terms of the last user's question



base-retriever	variant	zero-shot	CAsT'19 NDCG@3	CAsT'19 R@100	CAsT'20 NDCG@3	CAsT'20 R@100	CAsT'21 NDCG@3	CAsT'21 R@100
ColBERT	last-turn [a]	✔	0.214	0.157	0.155	0.124	0.140	0.154
	all-history [b]	✔	0.190	0.165	0.150	0.166	**0.237**	0.265
	ZeCo^2 *(ours)*	✔	0.238 [b]	**0.216** [a,b,c]	**0.176** [b]	**0.200** [a,b,c]	0.234 [a]	**0.267** [a]
	human		0.430	0.363	0.443	0.408	0.431	0.403
ConvDR [31]	zero-shot [c]	✔	**0.247**	0.183	0.150	0.150	–	–
	few-shot		0.466	0.362	0.340	0.345	0.361	0.376
	human		0.461	0.389	0.422	0.454	0.548	0.451

[Krasakis et al., 2022]

CAsT Year 3 Results



Conclusion

- Models evolved from **closed short answer** ConvQA models towards **open-retrieval models** that **generate short (and long)** answer responses
- Most effective methods have **pipelines** involving multiple components: **query rewriting**, query expansion, **dense retrieval**, **multi-pass re-ranking**, and result fusion.
- A common pattern is **stacking models** to add richer conversational modeling capability (Flow, 3D-CNNs, RGNNs).
- **Models**, **datasets**, and **evaluation** need to evolve to handle **richer forms of interactions** beyond questions and answers.