UNIVERSITY OF AMSTERDAM

A Comparison of Question Rewriting Methods for Conversational Passage Retrieval

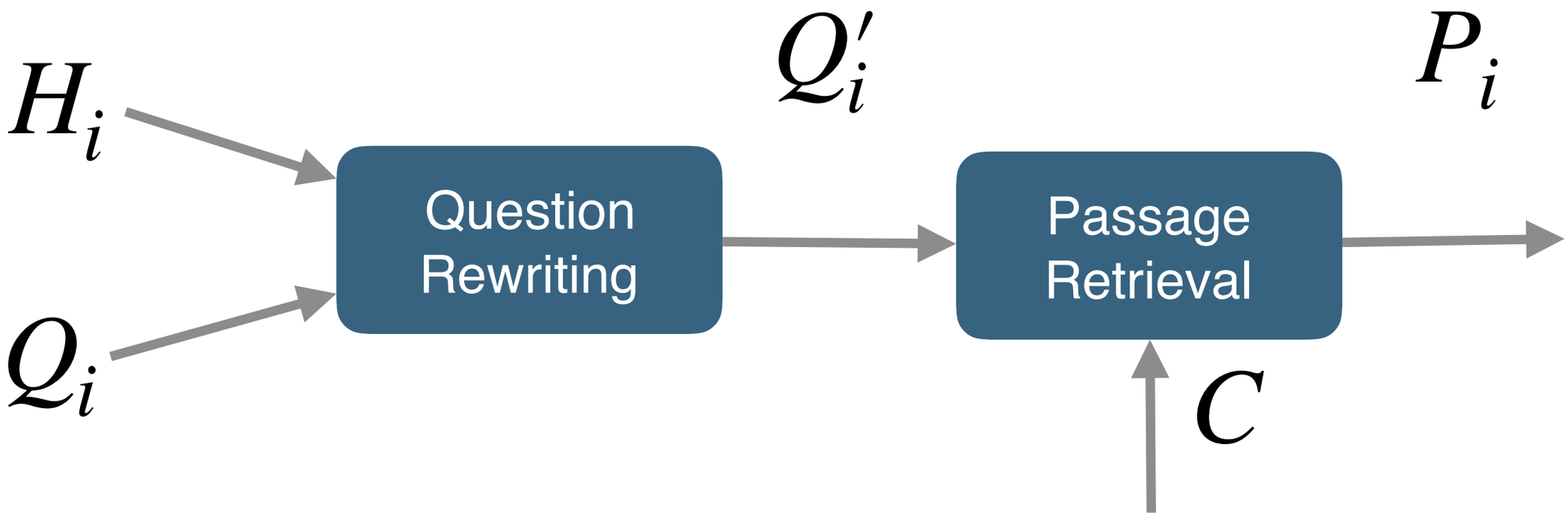
Svitlana Vakulenko, Nikos Voskarides, Zhucheng Tu, Shayne Longpre

UNIVERSITY OF AMSTERDAM

Conversational Passage Retrieval

Q1	What are some **interesting facts about bees**?	**Fun facts about bees**: 1 Honeybees are the only insect that produces food eaten by humans. 2 Worker honeybees are female. 3 The average worker bee produces only 1/12 teaspoon of honey over her lifetime. 4 A worker bee lives about 6 weeks. The queen bee can live to be 5 years old. 5 **Honey never spoils**.
Q2	Why doesn't **it** spoil?	Diana Taliun/iStock/Getty Images. **Honey doesn't spoil** like other foods and even if it has turned cloudy, it's still safe to eat. While you may be used to seeing clear, golden honey, cloudy honey is normal and natural.Clouding is caused by crystallization of the sugars and doesn't affect the taste or safety of the honey and is reversible.If the honey is in a jar, you can warm it on the stove. It doesn't harm honey to store it in the refrigerator, but it isn't necessary and will make honey cloudy and, sometimes, develop graininess as it crystallizes. 2 You can also warm honey in the microwave, as long as it is in a microwave-safe container and you use your microwave's lowest setting.
Q3	Why are so **many dying**?	The group found that **large numbers of bees are dying** during the summer months, when conditions should be more favorable. One in 4 colonies is now dying during summer, which was unheard of several years ago, according to the results. The total number of bee colonies in the United States declined from 6 million during the 1940s to 2.5 million about 10 years ago, but it has remained relatively stable since then.

Conversational Passage Retrieval



- rewritten question Q_i'
- passage collection C
- top-k relevant passages P_i

Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre: Question Rewriting for Conversational Question Answering. WSDM 2021.

Question Rewriting

✖ Supervised sequence generation

✖ Transformer++

Input: What are..? <SEP> .. Honey never spoils <SEP> Why doesn't it spoil?
Output: Why doesn't honey spoil?

- fine-tune GPT2
- CANARD 35K conversational questions + rewrites

Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre: Question Rewriting for Conversational Question Answering. WSDM 2021.

Question Rewriting

✖ **Weakly** supervised sequence generation

✖ Self-learn & Rule-based

Input: What are..? <SEP> .. Honey never spoils <SEP> Why doesn't it spoil?
Output: Why doesn't honey spoil?

- fine-tune GPT2
- **MS MARCO** 152K sessions -> conversations

Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett et al. Few-Shot Generative Conversational Query Rewriting. SIGIR 2020. **Best short paper award**.

Question Rewriting

✖ Supervised **sequence classification**

✖ QuReTeC

Input: What are..? <SEP> .. Honey never spoils <SEP> Why doesn't it spoil?
Output: 0 0 1 0 0
QR: Why doesn't it spoil? **Honey**

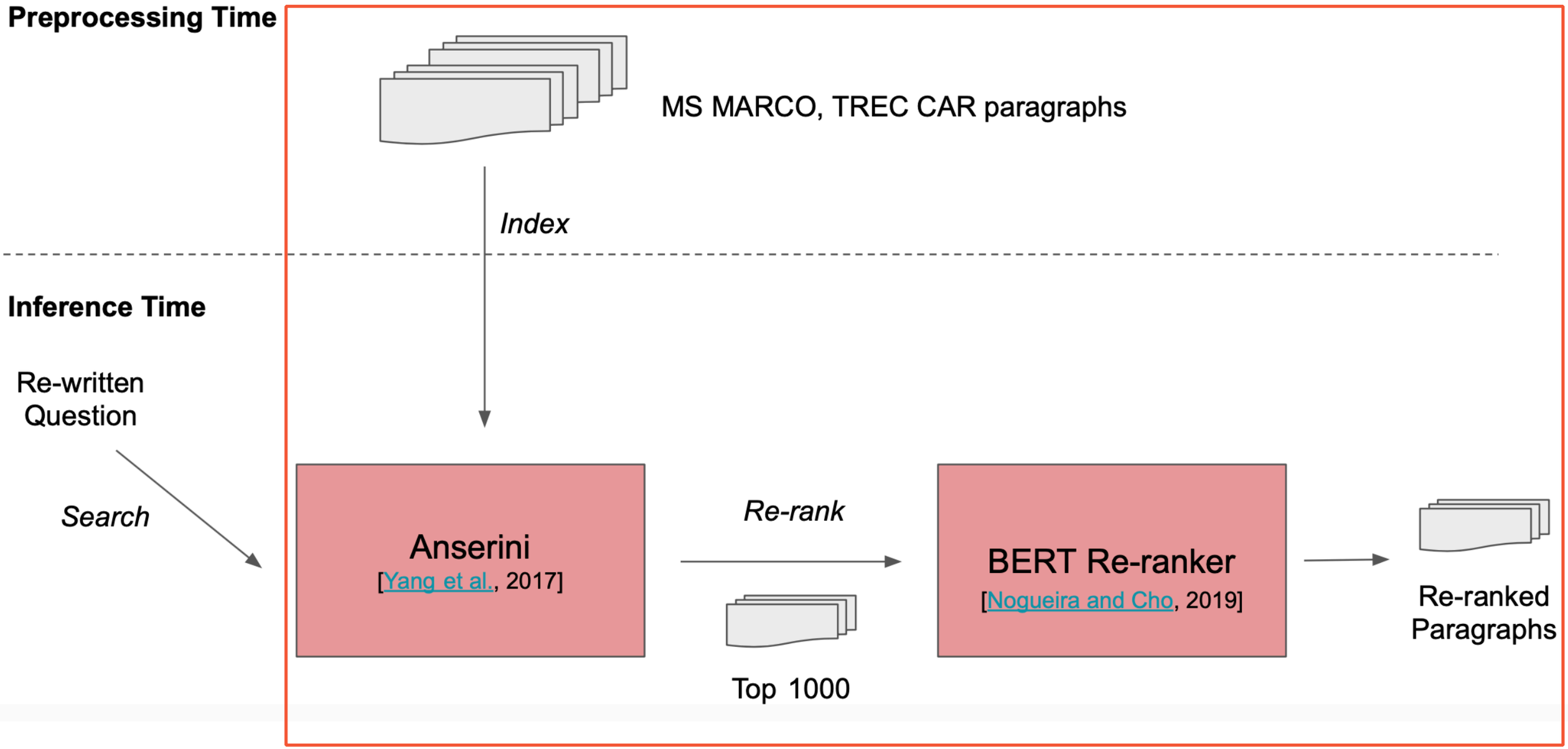
- fine-tune **BERT**
- CANARD 35K conversational questions + rewrites

Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, Maarten de Rijke: Query Resolution for Conversational Search with Limited Supervision. SIGIR 2020.

Research Questions

- **RQ1**: How do different QR approaches perform?

Conversational Passage Retrieval



Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre: Question Rewriting for Conversational Question Answering. WSDM 2021.

TREC CAsT Datasets

Dataset	#Topics	#Questions	#Copy (%)
CAsT 2019	50	479	88 (21)
CAsT 2020	25	216	5 (3)

- 2020
  - questions also depend on previous answer **Q&A**
  - canonical answers

Research Questions

- **RQ1**: How do different QR approaches perform?
- **RQ2**: Will their combination improve performance?

Results Fusion

✖ rewrite + terms predicted by QuReTeC

Transformer++: What do spanish people eat for dinner

QuReTeC: What do they eat for dinner? spanish christmas people

What do spanish people eat for dinner christmas

https://github.com/svakulenk0/cast_evaluation

Results: CAsT 2019

QR Method	Recall@1000	NDCG@3		ROUGE-1		
	Initial	Initial	Reranked	P	R	F
Original	0.417	0.131	0.266	0.92	0.76	0.82
Transformer++ Q	0.743	0.265	**0.525**	**0.96**	0.88	**0.91**
Self-Learn Q	0.725	0.261	0.513	0.93	0.89	0.90
Rule-Based Q	0.717	0.248	0.487	0.94	0.89	0.91
QuReTeC Q	**0.768**	**0.296**	0.500	0.89	**0.90**	0.89
Transformer++ Q + QuReTeC Q	**0.791**	0.300	**0.546**	**0.93**	0.91	**0.91**
Self-Learn Q + QuReTeC Q	0.785	0.293	0.519	0.90	**0.93**	**0.91**
Rule-Based Q + QuReTeC Q	0.783	**0.301**	0.534	0.91	**0.93**	**0.91**
Human-BoW Q	0.769	0.297	0.524	0.91	0.90	0.90
Human	0.803	0.309	0.577	1.00	1.00	1.00

Results: CAsT 2019

QR Method	Recall@1000	NDCG@3		ROUGE-1		
	Initial	Initial	Reranked	P	R	F
Original	0.417	0.131	0.266	0.92	0.76	0.82
Transformer++ Q	0.743	0.265	**0.525**	**0.96**	0.88	**0.91**
Self-Learn Q	0.725	0.261	0.513	0.93	0.89	0.90
Rule-Based Q	0.717	0.248	0.487	0.94	0.89	0.91
QuReTeC Q	**0.768**	**0.296**	0.500	0.89	**0.90**	0.89
Transformer++ Q + QuReTeC Q	**0.791**	0.300	**0.546**	**0.93**	0.91	**0.91**
Self-Learn Q + QuReTeC Q	0.785	0.293	0.519	0.90	**0.93**	**0.91**
Rule-Based Q + QuReTeC Q	0.783	**0.301**	0.534	0.91	**0.93**	**0.91**
Human-BoW Q	0.769	0.297	0.524	0.91	0.90	0.90
Human	0.803	0.309	0.577	1.00	1.00	1.00

Results: CAsT 2019

QR Method	Recall@1000	NDCG@3		ROUGE-1		
	Initial	Initial	Reranked	P	R	F
Original	0.417	0.131	0.266	0.92	0.76	0.82
Transformer++ Q	0.743	0.265	**0.525**	**0.96**	0.88	**0.91**
Self-Learn Q	0.725	0.261	0.513	0.93	0.89	0.90
Rule-Based Q	0.717	0.248	0.487	0.94	0.89	0.91
QuReTeC Q	**0.768**	**0.296**	0.500	0.89	**0.90**	0.89
Transformer++ Q + QuReTeC Q	**0.791**	0.300	**0.546**	**0.93**	0.91	**0.91**
Self-Learn Q + QuReTeC Q	0.785	0.293	0.519	0.90	**0.93**	**0.91**
Rule-Based Q + QuReTeC Q	0.783	**0.301**	0.534	0.91	**0.93**	**0.91**
Human-BoW Q	0.769	0.297	0.524	0.91	0.90	0.90
Human	0.803	0.309	0.577	1.00	1.00	1.00

UNIVERSITY OF AMSTERDAM

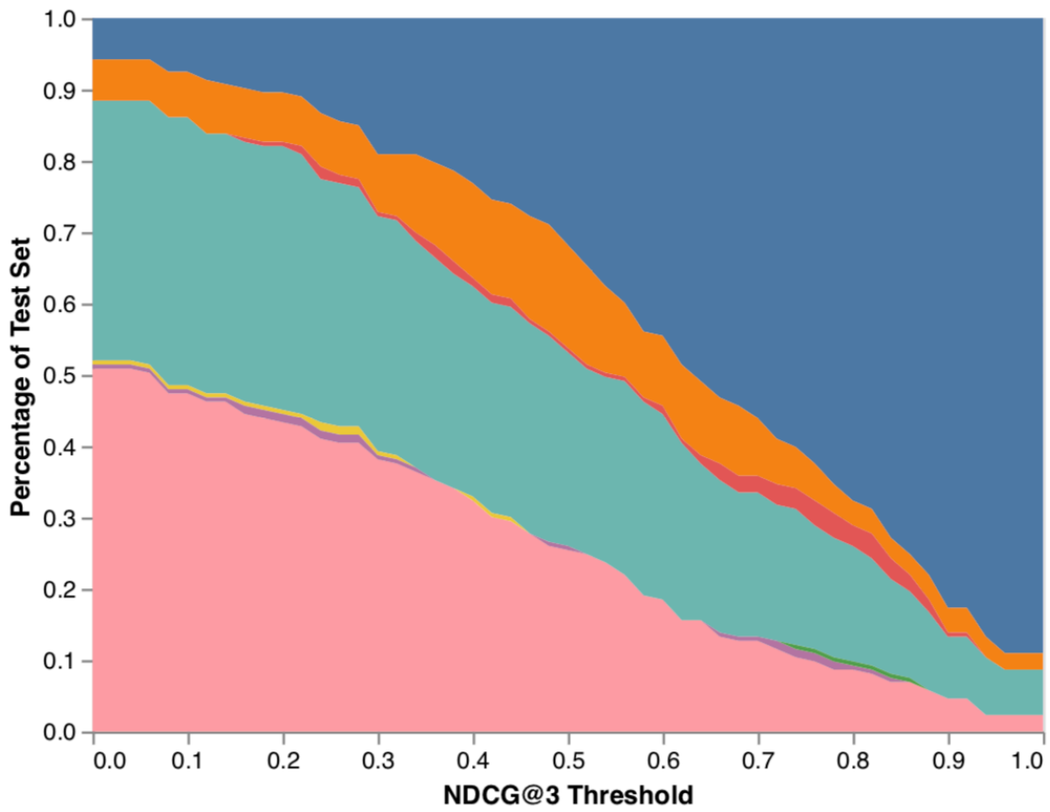
Results: CAsT 2020

QR Method	Recall@1000	NDCG@3		ROUGE-1		
	Initial	Initial	Reranked	P	R	F
Original	0.251	0.068	0.193	**0.87**	0.66	0.74
Transformer++ Q&A	0.351	0.098	0.252	0.75	0.69	0.70
Self-Learn Q&A	0.462	0.156	0.342	0.84	0.73	0.76
Rule-Based Q&A	0.455	0.137	0.339	0.84	0.75	**0.78**
QuReTeC Q&A	**0.531**	**0.171**	**0.370**	0.82	**0.77**	**0.78**
Transformer++ Q + QuReTeC Q&A	0.525	0.160	0.351	**0.83**	0.77	0.78
Self-Learn Q + QuReTeC Q&A	**0.567**	0.168	**0.375**	0.82	**0.79**	**0.79**
Rule-Based Q&A + QuReTeC Q&A	0.519	**0.173**	0.362	0.80	**0.79**	0.78
Human-BoW Q	0.579	0.189	0.465	0.89	0.81	0.84
Human-BoW Q&A	0.649	0.226	0.465	0.88	0.85	0.86
Human	0.707	0.240	0.531	1.00	1.00	1.00

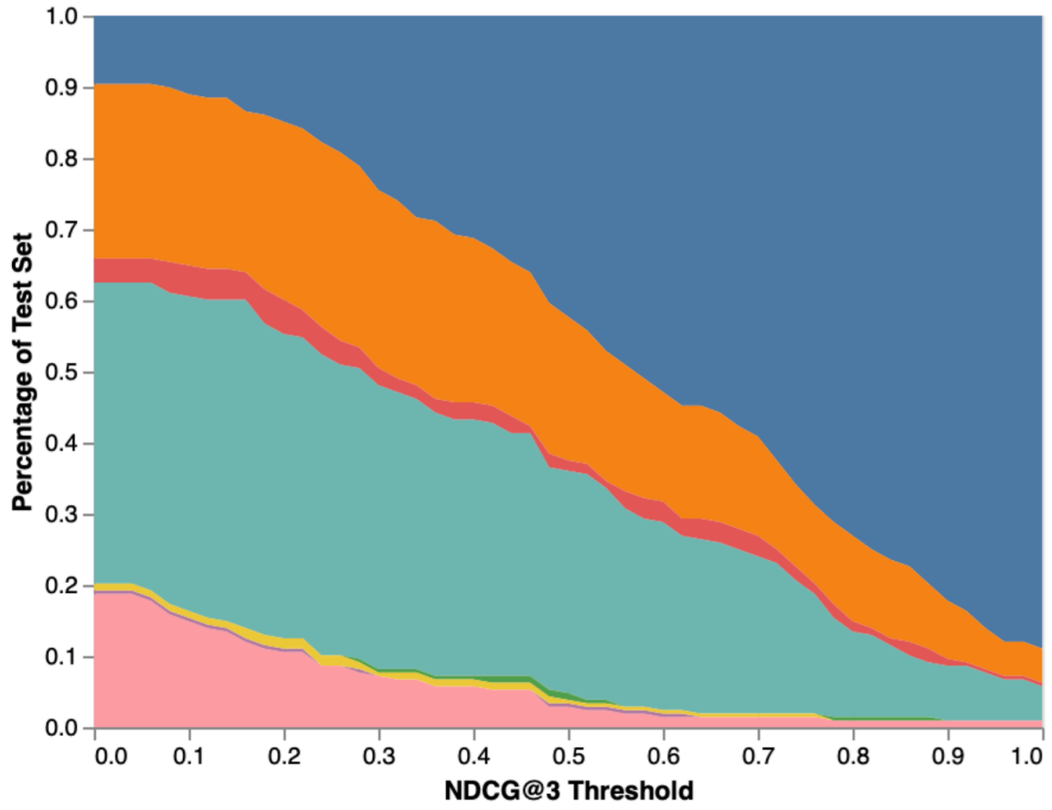
CAsT 2019 vs CAsT 2020

https://github.com/svakulenk0/QRQA

Transformer++ on CAsT 2019



QuReTeC on CAsT 2020



Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, Raviteja Anantha. A Wrong Answer or a Wrong Question? An Intricate Relationship between Question Reformulation and Answer Selection in Conversational Question Answering. SCAI@EMNLP 2020. **Best paper award**.

Svitlana Vakulenko, Nikos Voskarides, Zhucheng Tu, Shayne Longpre. Leveraging Query Resolution and Reading Comprehension for Conversational Passage Retrieval. TREC 2020.

Conclusion

✖ sequence classification for QR **improves recall**

✖ but it has a **lower upper bound** on performance

✖ **combining with generative models** helps

- more elegant fusion approaches for future work