Query Resolution for Conversational Search with Limited Supervision

Nikos Voskarides, Dan Li, Pengjie Ren,

Evangelos Kanoulas, Maarten de Rijke

Presentation by: Rafael Ferreira
Date:23/07/2020

Content

Multi-turn passage retrieval

- Let $[q_1, \ldots, q_{i-1}, q_i]$ be a sequence of conversational queries that share a common topic T. q_i is the current query and $q_{1:i-1}$ is the conversation history. Given q_i and $q_{1:i-1}$, the task is to retrieve a ranked list of passages L from a passage collection D that satisfy the user's information need.
- Multi-turn passage retrieval can be seen as a **instance of conversational search** in which each query takes into account the context from previous ones.
- **Challenges:**
  - Queries in conversational setting
  - Coreferences
  - Zero Anaphora (implicit coreferences)
  - Topic Shift

Conversational Search - Example


Q1

What is a **physician's assistant**?

Is a health care practitioner who practices medicine in collaboration with or under the (indirect) supervision of a physician.


A1


Q2

What are the educational requirements required to become **one**? *[coreference]*

In most cases, a physician assistant will need a master's degree from an accredited institution (two years of post-graduate education after completing a four-year degree).


A2


Q3

What is a **registered nurse**? *[topic shift]*

A registered nurse is a nurse who has graduated from a nursing program and met the requirements to obtain a nursing license.


A3


Q4

What is the difference between a **RN** and a **PA**? *[context needed to decipher PA and a NP]*

The RN model draws from the nursing tradition, including the whole person and wellness. The PA tradition draws more from a medical model.


A4

Manually Constructed Conversation

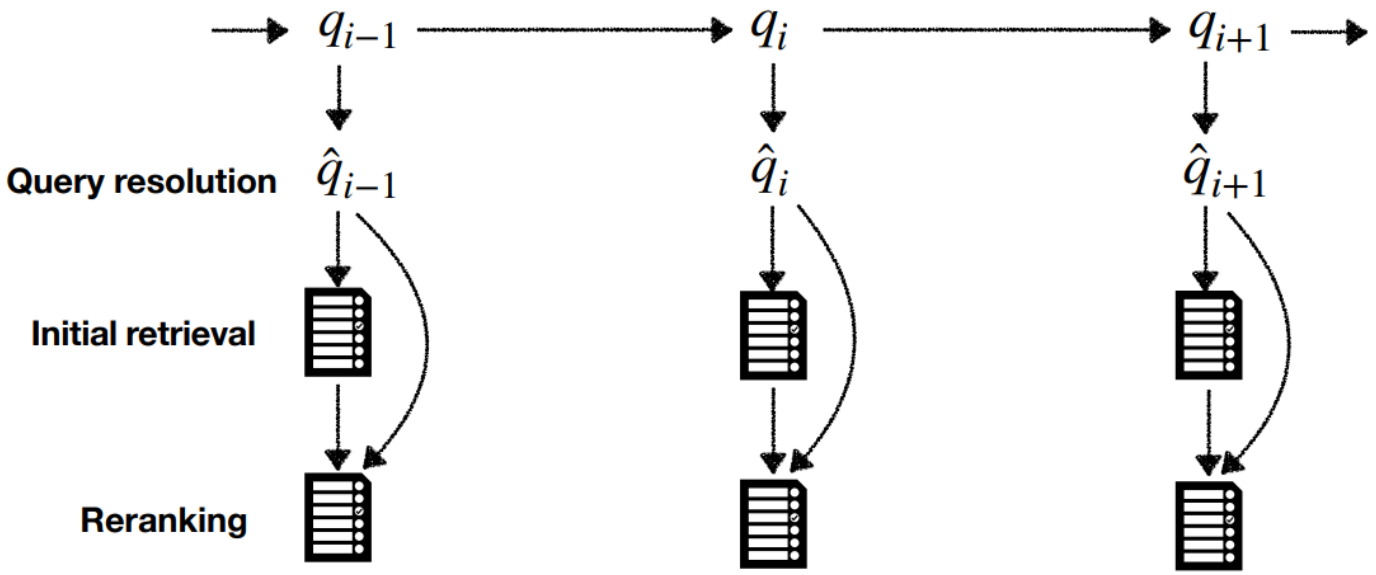
Paper Contributions

1. **QuReTeC**, a model for query resolution trained using a **binary term classification task** with the use of neural model based on **bidirectional transformers** (BERT[1]).
2. Proposed a **distant supervision method** that allows the use of general-purpose passage relevance data. This reduces the amount of human-curated data required to train a model.
3. Showed that using **QuReTeC** in a multi-stage ranking architecture is possible to **outperform other baseline models.**

[1]Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

Multi-Turn Passage Retrieval Pipeline (1)

1. **Query resolution** – Apply the query resolution algorithm resulting in $\hat{q}$ Initial ranking step.
2. **Traditional IR model** such as BM25, or LMD (Language model Dirichlet).

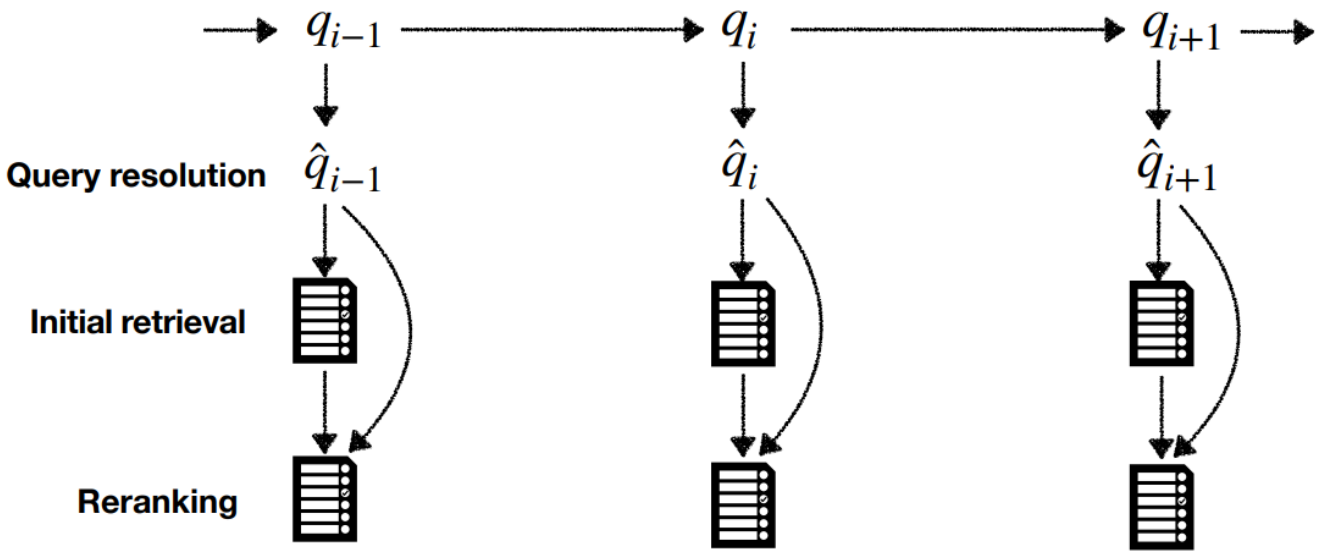


Nikos Voskarides et al. "Query Resolution for Conversational Search with Limited Supervision"

Multi-Turn Passage Retrieval Pipeline (2)

3. **Re-ranking step** - supervised neural ranker based on bidirectional transformers (**BERT**) with a **liner layer** with dropout on top.
   - In this work was used a BERT Base model finetuned on a **binary passage relevance classification task**[1], on the MS MARCO[2] dataset (contains 400 million tuples of query, relevant passage, non-relevant passage).

Input to BERT given in format: < [CLS] $\hat{q}_i$ [SEP] p >



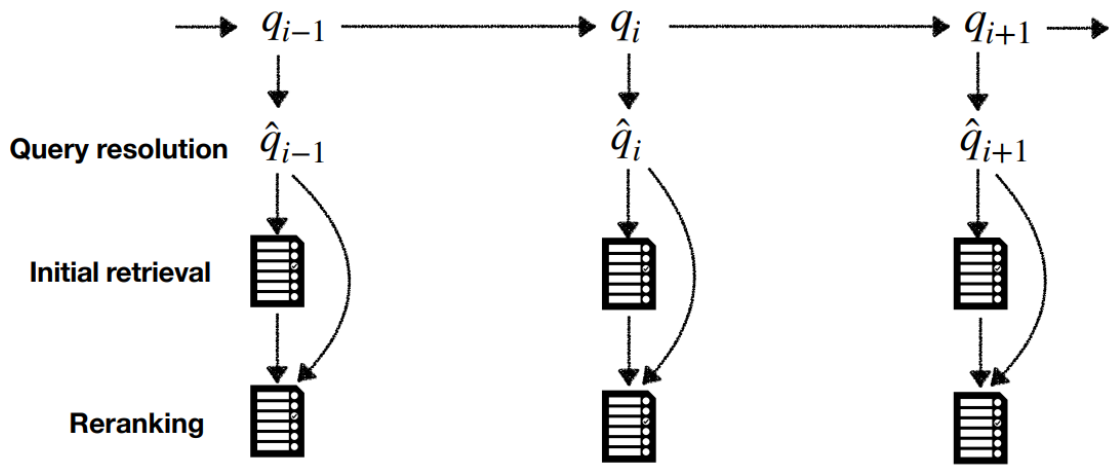
[1]Rodrigo Nogueira et al., "Passage Re-ranking with BERT"
[2]Payal Bajaj et al, "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset"
Nikos Voskarides et al. "Query Resolution for Conversational Search with Limited Supervision"

Multi-Turn Passage Retrieval Pipeline (3)

4. **Rank fusion** – method of combining the ranks from two different lists. In particular is used the **Reciprocal Rank Fusion**[1] (RRF) because of its simplicity (only one hyperparameter).

$$f_2(p) = \sum_{L' \in \{L_1, L_n\}} \frac{1}{k + rank(p, L')},$$

Where *rank(p, L')* is the rank of passage p in a ranked list L', and k is a hyperparameter.



[1]Gordon Cormack et al "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods"
Nikos Voskarides et al. "Query Resolution for Conversational Search with Limited Supervision"

Query Resolution - Method

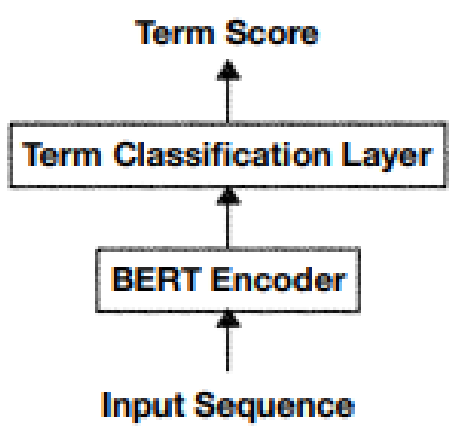
- Previous attempts use **sequence-to-sequence** models to rewrite queries.
  - Problems: needs large amounts of uncommon and cumbersome to obtain data.
- In this work the task is modeled as a **binary term classification task,** where given the history of the conversation and the current query, the output is a **binary label** (relevant or non-relevant) for each term in the history. Being the relevant terms then **concatenated** to the current query in order to resolve the context.
- **Relevant resolution terms** are defined by the expression:

$$E^*_{q_i} = terms(q^*_i) \cap terms(q_{1:i-1}) \setminus terms(q_i)$$

- Where q_i^* is the gold standard resolution for the current query. E^*_{qi} represents the context terms that are missing in the current query.

Nikos Voskarides et al. "Query Resolution for Conversational Search with Limited Supervision"

Query Resolution Model – QuReTeC (1)



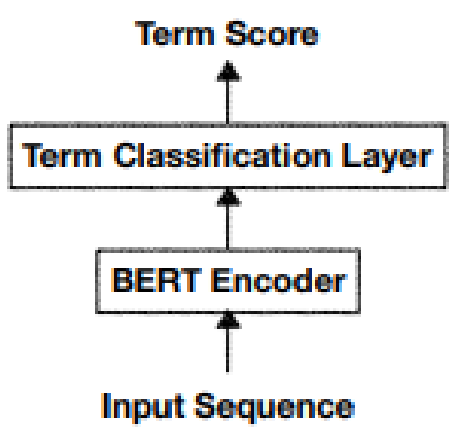
QuReTeC - **Qu**ery **Re**solution by **Te**rm **C**lassification

1. First is used **BERT** (in this work BERT Large) to obtain the embeddings for the history and queries. The output for the CLS token and current query are masked since we don't want to predict anything for those tokens.

Label	-	0	0	**1**	0	0	0	0	0	0	0	0	**1**	0	-	-	-	-	-	-
Input Sequence	<CLS>	*Who*	*formed*	*Saosin?*	*When*	*was*	*the*	*band*	*formed?*	*What*	*was*	*their*	*first*	*album?*	<SEP>	*When*	*was*	*the*	*album*	*released*
		Turn #1			Turn #2					Turn #3						Turn #4 (current)				

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

Query Resolution Model – QuReTeC (2)



QuReTeC - **Qu**ery **Re**solution by **Te**rm **C**lassification

2. Secondly is applied a **term classification layer** over the sub-tokens of each term. This layer is a simple linear layer with dropout and a sigmoid activation function.

- The loss function is the **binary cross-entropy**.

Nikos Voskarides et al. "Query Resolution for Conversational Search with Limited Supervision"

Query Resolution – Distant Supervision Labels

- Since the gold standard queries are not always available the authors create **distant supervision labels** in order to **increase the amount of training data**.
- The **gold answer are replaced by a relevant passage** to extract the relevant terms. By using the data in this way we have access to additional and more commonly available data.
- These distant supervision labels generate **noisy data,** but the rationale is that **increasing the amount of data outweighs this problem**.

Experimental Setup - Datasets

The model was evaluated on two distinct settings:

- **Extrinsic evaluation – Performance in retrieval and re-ranking:**
  - **TREC CAsT** dataset[1] (multi-turn passage retrieval)
  - Each topic is set of queries and the output should be a ranked list of passages.
  - Collection composed by MS MARCO[2] and TREC CAR[3] (passages from Wikipedia).
  - Annotation using a relevance scale from 0 to 4 in the test set.
- **Intrinsic evaluation – Performance in query resolution:**
  - **QuAC**[4] dataset (multi-turn span QA) – dialogues in a conversational format about a Wikipedia article..
  - **CANARD**[5] dataset (query rewriting dataset)– version of QuAC but the questions are rewritten in a non-conversational format.
  - Answer is span from the article.
  - Since these are not passage retrieval datasets, the authors used a 50-character window over the span given as answer, in order to capture the context.

[1]J. Dalton et al, "The TREC Conversational Assistance Track (CAsT)" - http://www.treccast.ai/.
[2]Payal Bajaj et al, "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset"
[3]Laura Dietz et al, "TREC Complex Answer Retrieval Overview"
[4]Eunsol Choi et al, "QuAC : Question Answering in Context"
[5]Ahmed Elgohary et al, "Can You Unpack That? Learning to Rewrite Questions-in-Context"

Experimental Setup - Metrics

- **Extrinsic evaluation – Performance in retrieval and re-ranking:**
  - **NDCG@3** - gives a penalty that is proportional to the relevance and position of the result in the top 3 documents retrieved.
  - **Recall@1000** - fraction of documents that are relevant to the query that were retrieved successfully.
  - **MAP@1000** - Mean Average Precision is for a set of queries, the mean of the average precision scores for each query.
  - **MRR@1000** - Mean Reciprocal Rank is used to calculate the reciprocal of the rank at which the first relevant document was retrieved.

- **Intrinsic evaluation – Performance in query resolution**
  - **Micro-Precision**
  - **Micro-Recall**
  - **Micro-F1**
  - Metrics calculated by query averaging over all turns and topics.

Experimental Setup – Baselines (1)

- **Intrinsic and extrinsic baselines:**
  - **Original** – Using the original conversational queries (with/without concatenation of queries from other turns).
  - **RM3**[1] – Pseudo relevance feedback model.
  - **Neural Coref**[2] – A coreference resolution method designed for chatbots. It uses a rule-based system for mention detection and a feed-forward neural network that predicts coreference scores.
  - **BiLSTM-copy**[3] – A neural sequence-to-sequence model for query resolution. It uses a BiLSTM encoder and decoder augmented with attention and copy mechanisms.

[1]Nasreen Abdul-Jaleel et al, UMass at TREC 2004: Novelty and HARD
[2]https://medium.com/huggingface/state-of-the-art-neural-coreference-resolution-for-chatbots-3302365dcf30
[3]Ahmed Elgohary et al, "Can You Unpack That? Learning to Rewrite Questions-in-Context"

Experimental Setup – Baselines (2)

- **Only extrinsic baselines – Initial Retrieval:**
  - **Nugget**[1] - Extracts substrings from the current and previous turn queries to build a new query for the current turn.
  - **QCM**[2] - Models the edits between consecutive queries and the results list returned by the previous turn query to construct a new query for the current turn.
  - **Oracle** - Performs initial retrieval using the gold standard resolution query.

- **Only extrinsic baselines – Re-ranking:**
  - **TREC-top-auto** - Uses an automatic system for query resolution and BERT-large for reranking.
  - **TREC-top-manual** - Uses the gold standard query resolution and BERT-large for reranking.

[1]Dongyi Guan et al, "Effective Structured Query Formulation for Session Search"
[2]Hui Yang et al, "The Query Change Model: Modeling Session Search as a Markov Decision Process"

Results Intrinsic Evaluation – Query resolution (1)

Method	P	R	F1
Original (cur+prev)	22.3	46.4	30.1
Original (cur+first)	41.1	49.5	44.9
Original (all)	12.3	**100.0**	21.9
NeuralCoref	65.5	30.0	41.2
BiLSTM-copy	67.0	53.2	59.3
QuReTeC	**71.5**	66.1	**68.7**

Intrinsic evaluation for query resolution on the QuAC test set

- In **QuAC QuReTeC improves in most metrics** when compared to original and neural baselines.

Results Intrinsic Evaluation – Query resolution (2)

Method	P	R	F1
Original (cur+prev)	32.5	43.9	37.4
Original (cur+first)	43.0	74.0	54.4
Original (all)	18.6	**100.0**	31.4
RM3 (cur)	35.8	8.3	13.5
RM3 (cur+prev)	34.6	32.5	33.5
RM3 (cur+first)	40.9	32.9	36.5
RM3 (all)	41.5	38.8	40.1
NeuralCoref	**83.0**	28.7	42.7
BiLSTM-copy	51.5	36.0	42.4
QuReTeC	77.2	79.9	**78.5**

Intrinsic evaluation for query resolution on the TREC CAsT test set

- **CAsT** has similar results, showing the **capability of generalizing to this dataset**, since it was only trained on the QuAC dataset.

Results Extrinsic Evaluation – Initial Retrieval

- **Initial retrieval performance** – QuReTeC and simple retrieval methods (no reranking).
- **QuReTeC** is better than most baselines.
- By seeing the results of **Oracle** we are able to conclude that there is still **space for improvement**.

Method	Recall	MAP	MRR	NDCG@3
Original (cur)	0.438	0.129	0.310	0.155
Original (cur+prev)	0.572	0.181	0.475	0.235
Original (cur+first)	0.655	0.214	0.561	0.282
Original (all)	0.694	0.190	0.552	0.256
RM3 (cur)	0.440	0.140	0.320	0.158
RM3 (cur+prev)	0.575	0.200	0.482	0.254
RM3 (cur+first)	0.656	0.225	0.551	0.300
RM3 (all)	0.666	0.195	0.544	0.266
Nugget	0.426	0.101	0.334	0.145
QCM	0.392	0.091	0.317	0.127
NeuralCoref	0.565	0.176	0.423	0.212
BiLSTM-copy	0.552	0.171	0.403	0.205
QuReTeC	**0.754**▲	**0.272**▲	**0.637**▲	**0.341**▲
Oracle	0.785	0.309	0.660	0.361

Initial retrieval performance on the TREC CAsT test set for different query resolution methods

Results Extrinsic Evaluation – Re-ranking

- **Re-ranking performance** – re-ranking results using BERT after the initial retrieval.

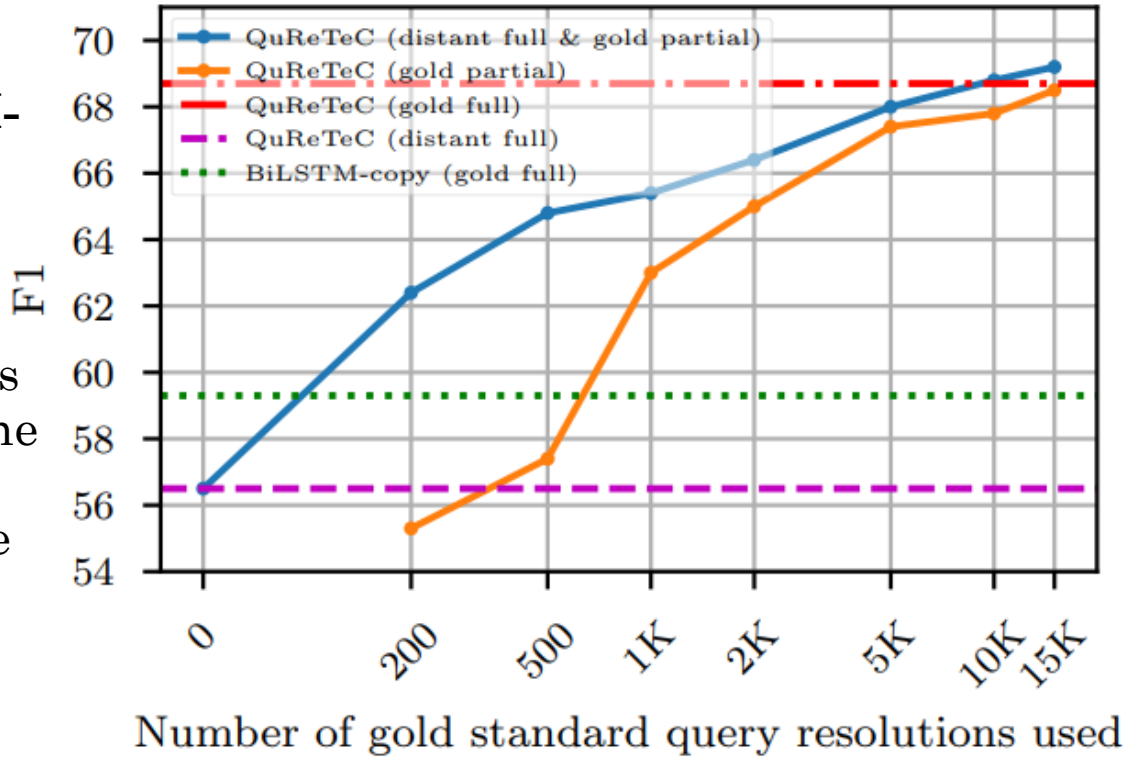
Method	MAP	MRR	NDCG@3
Initial	0.272	0.637	0.341
BERT-base	0.272	0.693	0.408
RRF (Initial + BERT-base)	**0.355**▲	**0.787**▲	**0.476**▲
Oracle	0.754	0.956	0.926
TREC-top-auto	0.267	0.715	0.436
TREC-top-manual	0.405	0.879	0.589

Re-ranking performance on the TREC CAsT test set. First group uses QuReTeC.

- The **re-ranking step improves** the results (as expected).
- **QuReTeC** with re-ranking achieves results close to the best reported in TREC CAsT.
- With the **addition of RRF, QuReTEC** is able to beat TREC-top-auto.
- From the Oracle and TREC-top-manual we can see that this is still an open problem.

Results - Distant Supervision for Query Resolution

- **QuReTeC is competitive with BiLSTM-copy** even without using gold-resolutions (distant full), and when using less data.
- Increasing the amount of data we see that the effect of using distant supervision labels when compared to gold labels is smaller. The **combination of both distant supervision and gold labels** achieves the best results.
- These results show that is possible to use distant supervision labels to **reduce the number of human-curated data needed** in scenarios where the amount of data available is small.



Query resolution performance (intrinsic) on the QuAC test set on different supervision settings. Gold refers to the QuAC train (gold) dataset and distant refers to the QuAC train (distant) dataset. Full refers to the whole and partial refers to a part of the corresponding dataset (gold or distant).

Nikos Voskarides et al. "Query Resolution for Conversational Search with Limited Supervision"

Conclusion

- In his paper is studied the task of **query resolution for conversational search**.
- The query resolution model proposed **QuReTeC**, uses a **bidirectional transformer architecture,** and is trained on a **binary term classification task**. It was shown that this model is very **useful in both rewriting and retrieval-ranking**.
- The use of **distant supervision labels** also proved to **improve the model when combined with the gold standard answers**.

Research Opportunities

- The authors propose the idea of creating **specialized rankers** for initial retrieval and ranking.
- In my opinion is important to study of **other transformer models** and how to handle **the context when it is too long.**
- **Entity** recognition, extraction, and linking.

Thank You!

Contacts:

- rah.ferreira@campus.fct.unl.pt
- randreferreira19@gmail.com