Table 5: Intrinsic evaluation for query resolution on the QuAC test set. Cur, prev, first and all refer to using the current, previous, first or all turns respectively.

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**

Table 6: Intrinsic evaluation for query resolution on the TREC CAsT test set. Cur, prev, first and all refer to using the current, previous, first, or all turns respectively.

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**

Baselines For RM3, we tune the following parameters: $n \in \{3, 5, 10, 20, 30\}$ and $k \in \{5, 10\}$ and set the original query weight to the default value of 0.8. For Nugget, we set $k_{snippet} = 10$ and tune $\theta \in \{0.95, 0.97, 0.99\}$. For QCM, we tune $\alpha \in \{1.0, 2.2, 3.0\}$, $\beta \in \{1.6, 1.8, 2.0\}$, $\epsilon \in \{0.06, 0.07, 0.08\}$ and $\delta \in \{0.2, 0.4, 0.6\}$. For both Nugget and QCM we use Van Gysel et al. [42]'s implementation. For fair comparison, we retrieve over the whole collection rather than just reranking the top-1000 results. The aforementioned methods are tuned on the small annotated training set of TREC CAsT. For query resolution, we tune the greedyness parameter of NeuralCoref in the range $\{0.5, 0.75\}$. We use the model of BiLSTM-copy released by [15], as it was optimized specifically for QuAC with gold standard resolutions.

Preprocessing We apply lowercase, lemmatization and stopword removal to q_i^*, $q_{1:i-1}$ and q_i using Spacy[12] before calculating term overlap in Equation 2.

6 RESULTS & DISCUSSION

In this section we present and discuss our experimental results.

6.1 Query resolution for multi-turn retrieval

In this subsection we answer (RQ1): we study how QuReTeC performs compared to other state-of-the-art methods when evaluated on term classification (Section 6.1.1), when incorporated in the initial retrieval step (Section 6.1.2) and in the reranking step (Section 6.1.3).

Table 7: Initial retrieval performance on the TREC CAsT test set for different query resolution methods. The retrieval model is fixed (same as in Section 3.2.1). Significance is tested against RM3 (cur+first) since it has the best NDCG@3 among the baselines.

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

6.1.1 *Intrinsic evaluation.*

In this experiment we evaluate query resolution as a term classification task.[13] Table 5 shows the query resolution results on the QuAC dataset. We observe that QuReTeC outperforms all the variations of Original and the NeuralCoref by a large margin in terms of F1, precision and recall – except for Original (all) that has perfect recall but at the cost of very poor precision. Also, QuReTeC substantially outperforms BiLSTM-copy on all metrics. Note that BiLSTM-copy was optimized on the same training set as QuReTeC (see Section 5.5). This shows that QuReTeC is more effective in finding missing contextual information from previous turns.

Table 6 shows the query resolution results on the CAsT dataset. Generally, we observe similar patterns in terms of overall performance as in Table 5. Interestingly, we observe that QuReTeC generalizes very well to the CAsT dataset (even though it was only trained on QuAC) and outperforms all the baselines in terms of F1 by a large margin. In contrast, BiLSTM-copy fails to generalize and performs worse than Original (cur+first) in terms of F1. NeuralCoref has higher precision but much lower recall compared to QuReTeC. Finally, RM3 has relatively poor query resolution performance. This indicates that pseudo-relevance feedback is not suitable for the task of query resolution.

6.1.2 *Query resolution for initial retrieval.*

In this experiment, we evaluate query resolution when incorporated in the initial retrieval step (Section 3.2.1). We compare QuReTeC to the baseline methods in terms of initial retrieval performance. Table 7 shows the results. First, we observe that QuReTeC outperforms all the baselines by a large margin on all metrics. Also, interestingly, QuReTeC achieves performance close to the one achieved by the Oracle performance (gold standard resolutions). Note that there is still plenty of room for improvement even when using Oracle, which indicates that

[12] http://spacy.io/

[13] Note that the performance of Original (cur) is zero by definition when using the current turn only (see Eq. 2). Thus, we do not include it in Tables 5 and 6. Also, RM3 is not applicable in Table 5 since QuAC is not a retrieval dataset.

Table 8: Reranking performance on the TREC CAsT test set. All the methods in the first group use QuReTeC for query resolution. Significance is tested against BERT-base.

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

exploring other ranking functions for initial retrieval is a promising direction for future work. QuReTeC outperforms all Original and RM3 variations, which perform similarly. The session search methods (Nugget and QCM) perform poorly even compared to the Original variations, which indicates that session search is different in nature than conversational search. BiLSTM-copy performs poorly compared to QuReTeC but also compared to the Original variations, which means that it does not generalize well to CAsT.

6.1.3 Query resolution for reranking. In this experiment, we study the effect of QuReTeC when incorporated in the reranking step (Section 3.2.2). We keep the initial ranker fixed for all QuReTeC models. Table 8 shows the results. First, we see that BERT-base improves over the initial retrieval model that uses QuReTeC for query resolution on the top positions (second line). Second, when we fuse the ranked listed retrieved by BERT-base and the ranked list retrieval by the initial retrieval ranker using RRF, we significantly outperform BERT-base on all metrics (third line). This shows that the two rankers can be effectively combined with RRF, which is a very simple fusion method that only has one parameter which we do not tune. We also see that our best model outperforms TREC-top-auto on all metrics. Furthermore, by comparing RRF (line 3) to Oracle (line 4) we see that there is still plenty of room for improvement for reranking, which is a clear direction for future work. This also shows that the TREC CAsT dataset is sufficiently challenging for comparing automatic systems. Note that TREC-top-manual uses the gold standard query resolutions and is thereby not directly comparable with the rest of the methods.

6.2 Distant supervision for query resolution

In this section we answer (RQ2): Can we use distant supervision to reduce the amount of human-curated query resolution data required to train QuReTeC? Figure 3 shows the query resolution performance when training QuReTeC under different settings (see figure caption for a more detailed description). For QuReTeC (distant full & gold partial) we first pretrain QuReTeC on distant and then resume training with different fractions of gold. First, we see that QuReTeC performs competitively with BiLSTM-copy even when it does not use any gold resolutions (distant full).[14] More importantly, when only trained on distant, QuReTeC performs remarkably well in the low data regime. In fact, it outperforms BiLSTM-copy (trained on gold) even when using a surprisingly low number of gold standard query resolutions (200, which is ~1% of gold). Last, we see that as we add more labelled data, the effect of distant supervision becomes smaller. This is expected and is also the case for the model trained on QuAC train (gold).[15]

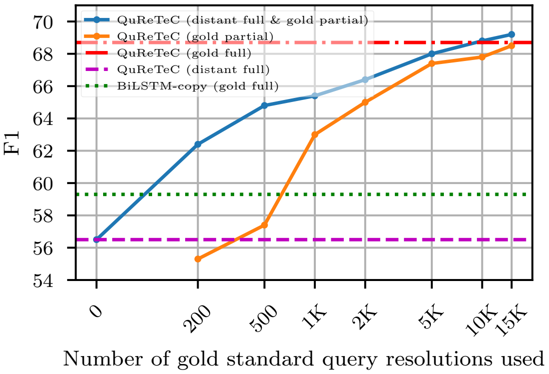
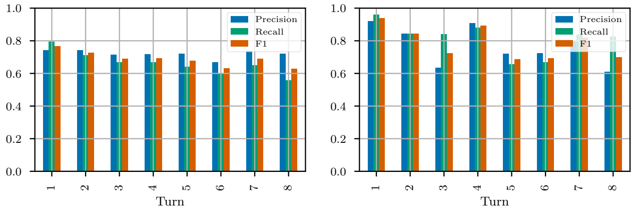


Figure 3: Query resolution performance (intrinsic) on the QuAC test set on different supervision settings. Gold refers to the QuAC train (gold) dataset and distant refers to the QuAC train (distant) dataset. Full refers to the whole and partial refers to a part of the corresponding dataset (gold or distant). The x-axis is plotted in log-scale.



(a) QuAC *test* **(b) CAsT *test***

Figure 4: Intrinsic query resolution evaluation (term classification performance) for QuReTeC, averaged per turn.

In order to test whether our distant supervision method can be applied on different encoders, we performed an additional experiment where we replaced BERT with a simple BiLSTM as the encoder in QuReTeC. Similarly to the previous experiment, we observed a substantial increase in F1 when retraining with 2K gold standard resolutions (+12 F1) over when only using gold resolutions.

In conclusion, our distant supervision method can be used to substantially decrease the amount of human-curated training data required to train QuReTeC. This is especially important in low resource scenarios (e.g. new domains or languages), where large-scale human-curated training data might not be readily available.

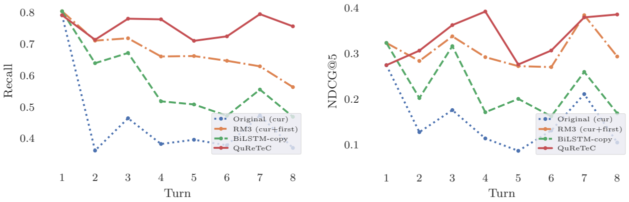
6.3 Analysis

In this section we perform analysis on QuReTeC when trained with gold standard supervision.

6.3.1 Query resolution performance per turn. Here we answer (RQ3) by analyzing the robustness of QuReTeC at later conversation turns.

[14] Also, when trained with distant full, QuReTeC performs better than an artificial method that uses the label of the distant supervision signal as the prediction in terms of F1 (56.5 vs 41.6). This is in line with previous work that successfully uses noisy supervision signals for retrieval tasks [12, 44].

[15] In fact (not shown in Figure 3), performance stabilizes after 15K query resolutions (~75% of gold full).



(a) Recall

(b) NDCG@5

Figure 5: Initial retrieval performance per turn for different query resolution methods CAsT *test*

Table 9: Qualitative analysis for QuReTeC on query resolution (intrinsic). We denote true positive terms with underline and false negative terms in italics. The examples are sampled from the QuAC dev set.

Success case – no mistakes

Q1: What was Bipasha Basu's debut?

A1: In 2001, Basu finally made her debut opposite Akshay Kumar in Vijay Galani 's Ajnabee.

Q2: Did this help her become well known?

A2: It was a moderate box-office success and attracted unfavorable reviews from critics.

Q3 (current): Why did she receive unfavorable reviews?

Failure case – misses two relevant terms: *dehusking*, *machine*

Q1: How old was Alexander Graham Bell when he made his first invention?

A1: The age of 12.

Q2: What did he invent?

A2: Bell built a homemade device that combined rotating paddles with sets of nail brushes.

Q3: What was it for?

A3: A simple *dehusking machine.*

Q4 (current): By inventing this, what happened to allow him to continue inventing things?

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.

We expect query resolution to become more challenging as the conversation history becomes larger (later in the conversation).

Intrinsic Figure 4 shows the QuReTeC performance averaged per turn on the QuAC and CAsT datasets. Even though performance decreases towards later turns as expected, we observe that it decreases very gradually, and thus we can conclude that QuReTeC is relatively robust across turns.

Extrinsic – initial retrieval Figure 5 shows the performance of different query resolution methods when incorporated in the initial retrieval step. We observe that QuReTeC is robust to later turns in the conversation, whereas the performance of all the baseline models decreases faster (especially in terms of recall). For reranking, we observe similar patterns as with initial retrieval; we do not include those results for brevity.

6.3.2 Qualitative analysis. Here we perform qualitative analysis by sampling specific instances from the data.

Intrinsic Table 9 shows one success and one failure case for QuReTeC from the QuAC dev set. In the success case (top) we observe that QuReTeC succeeds in resolving "she" ⟶ {"Bipasha", "Basu"} and "reviews" ⟶ "Anjabee". Note that "Anjabee" is a movie in which Basu acted but is not mentioned explicitly in the current turn. In the failure case (bottom) we observe that QuReTeC succeeds in resolving "him" ⟶ {"Alexander", "Graham" "Bell"} but misses the connection between "this" and "dehusking machine".

Extrinsic – initial retrieval Table 10 shows an example from the CAsT test set where QuReTeC succeeds and RM3 (cur+first), the best performing baseline for initial retrieval, fails. First, note that a topic change happens at Q7 (the topic changes from general real-time databases to Firebase DB). We observe that QuReTeC predicts the correct terms, and a relevant passage is retrieved at the top position. In contrast, RM3 (cur+first) fails to detect this topic change and therefore an irrelevant passage is retrieved at the top position that is about real-time databases on mobile apps but not about Firebase DB.

7 CONCLUSION

In this paper, we studied the task of query resolution for conversational search. We proposed to model query resolution as a binary term classification task: whether to add terms from the conversation history to the current turn query. We proposed QuReTeC, a neural query resolution model based on bidirectional transformers. We proposed a distant supervision method to gather training data for QuReTeC. We found that QuReTeC significantly outperforms multiple baselines of different nature and is robust across conversation turns. Also, we found that our distant supervision method can substantially reduce the required amount of gold standard query resolutions required for training QuReTeC, using only query-passage relevance labels. This result is especially important in low resource scenarios, where gold standard query resolutions might not be readily available.

As for future work, we aim to develop specialized rankers for both the initial retrieval and the reranking steps that incorporate QuReTeC in a more sophisticated way. Also, we want to study how to effectively combine QuReTeC with text generation query resolution methods as well as pseudo-relevance feedback methods. Finally, we aim to explore weak supervision signals for training QuReTeC [12].