TREC CAR Y3: Complex Answer Retrieval Overview

Laura Dietz[*], John Foley

Homepage: `http://trec-car.cs.unh.edu`
Google group mailinglist: `TREC-CAR`

1 Introduction

The vision of TREC Complex Answer Retrieval is to create complex long-form answers in response to a wide-variety information needs. In general, we aspire to create answers that are reminiscent to Wikipedia articles or school text books (e.g. TQA). However, while the vast majority of Wikipedia articles are about people, in TREC CAR we aim at information needs that are off the beaten path, covering topics in popular science, technology, and illnesses.

The first two years of TREC CAR, we aimed to reproduce Wikipedia articles. This provided a very large-scale automated benchmark, which had significant impact on neural ranking research [7, 6], as well as feature-based ranking models [1, 4]. The downside was that we had to prohibit access to Wikipedia, collections that could include Wikipedia (e.g. ClueWeb), and knowledge bases derived from Wikipedia (we provided the part of Wikipedia from which a knowledge graph can be built that excludes the benchmark topics, called "allButBenchmark"). Technically this would even affect resources that are trained on Wikipedia, such as most word embeddings and BERT [3]. To avoid this difficulty, in this year, our test topics come exclusively from an collection of school text book chapters, which are provided along with the TQA dataset [5]. These chapters have a similar length as Wikipedia articles, but are written for a younger audience. We derive outlines from TQA chapters, and ask participants to populate these outlines with paragraphs from Wikipedia, using the paragraphCorpus from previous years. We manually cleaned and rewrote the outlines so that they are suitable to be treated like search queries. This test collection is called benchmarkY3test.

A downside of this decision is that no automatic evaluation can be conducted. We recommend to train data-hungry methods on the "train" collection provided in the first year (Y1). Since previous year's test data (benchmarkY2test) contained both contained Wikipedia topics and TQA topics, we re-released manually assessed TQA topics as training data for this year, released as benchmarkY3train.

2 Worked Example

To motivate a brief example, consider a user interested in learning about water pollution through fertilizers, ocean acidification, and aquatic debris and the effects it has. There is no short and simple answer to this information need. Instead we need to retrieve a complex answer that covers the topic with its different facets and elaborates pertinent connections between entities/concepts. A suitable answer would cover the following:

> Through photosynthesis algae provide food and nutrients for the marine ecosystem. However, through rain storms, fertilizers used in agriculture and lawn care are swept into the rivers and coastal sea. Fertilizers contain nitrogen and phosphorys, these stimulate algae growth so that the alae population will grow large very quickly, called algal blooms. The problem is that these algae do not live long, when they die and decompose oxygen is removed from the water. As a result fish and shell fish die.

[*] dietz@cs.unh.edu

> Furthermore, some algal blooms release toxins into the water, which are consumed by shell fish. Humans that consume toxins though shell fish can suffer neurological damage.
>
> A different source of water pollution is through high levels of carbon dioxide in the athmosphere. Oceans absorb carbon dioxide, but it will lower the PH level of the water, meaning that oceans to become acidic. As a result, corals and shell fish are killed and other marine organisms reproduce less. This leads to issues in the food chain, and thereby less fish and shell fish for humans to consume.
>
> Finally, trash and other debris that gets in the waterways through shipping accidents, landfill erosion, or by directly dumping trash in the ocean. This debris is dangerous for aquatic wildlife in two ways. Animals may mistake debris for food swallow plastic bags which kills them. Other aquatic animals are tangled in nets and strangled by trash like plastic six-pack rings.

This example was taken from the TQA collection, "Effects of Water Pollution", and many similar examples can be found on Wikipedia. Nevertheless, such articles are not available for any imaginable kind of of information needs, which is why we aim to generate such comprehensive summaries automatically from Web sources through passage retrieval, consolidation, and organization. Of course, one might envision other responses that would satisfy the information need equally well.

3 Task Description

While in the first two years, the task setup followed a standard ad hoc IR fashion, where given a title and an outline of headings, a ranking of paragraphs is to be produced for every section. However, the long-term vision of CAR is to produce comprehensive articles. Articles are much different from rankings: Instead of ordering paragraphs by relevance, paragraphs should be ordered so that when read from top to bottom would would make logical sense in order to inform the user.

Y3 Passage Ordering Task Given an outline Q, retrieve, select, and arrange a sequence of k passages P from the provided passage corpus, with ideally:

1. High relevance of all passages
2. Balanced coverage of all query facets as defined through headings H_i in the outline
3. Maximizing topical coherence, minimizing topic switches, i.e., first all passages about one topic, then all passages of the next topic while avoiding to interleave multiple topics.

The number of passages k is given with the topic.

We are aware that this is a major departure from ad hoc retrieval. To facilitate the transition, organizers provided a script that would take the top ranked passages for Y1/Y2-style task (one ranking for each title+heading query), and would use them to populate the article (in order of relevance) so that exactly k passages are contained.

4 Topic Coordination with CAsT

Since the nature of topics for TREC CAR is similar to a subset of topics in TREC Conversational Assistance (CAsT)[2], both track's organizers coordinated the set of topics for assessment to prefer similar domains. We hope that this leads to interesting further research improving information access as (non-interactive) long-form response as well as a conversation. See Table 1 for a list of CAR and CAsT topics that share the topic.

Table 1: Related CAR and CAsT topics (many more CAR topics were assessed).

	CAsT	CAR
Cancer and non-infectuous diseases	cast2019:31	tqa2:L_0402
Sharks	cast2019:32	tqa2:L_0474
Lyme disease	cast2019:38	tqa2:L_0502, tqa2:L_0398
Satellites and space	cast2019:50	tqa2:L_0040, tqa2:L_0051, tqa2:L_0052
Evolution	cast2019:56	tqa2:L_0432
Injuries	cast2019:59	tqa2:L_0385, tqa2:L_0398
Blood cells	cast2019:67	tqa2:L_0402, tqa2:L_0385
Solar energy	cast2019:70	tqa2:L_0074, tqa2:L_311
Diet and health	cast2019:78	tqa2:L_0402

5 Assessment of the Manual Ground Truth

For 55 topics, we fully assessed three runs (i.e., generated articles) from each team. To study inter-annotator agreement, we selected three runs, and asked all six judges to annotate a separate topic (tqa2:L_0257) for each of these runs.

5.1 Assessment Interface

A screenshot of the assessment interface is given in Figure 1.

After reading the gold article for the query the judges were asked to assess the page from top to bottom by

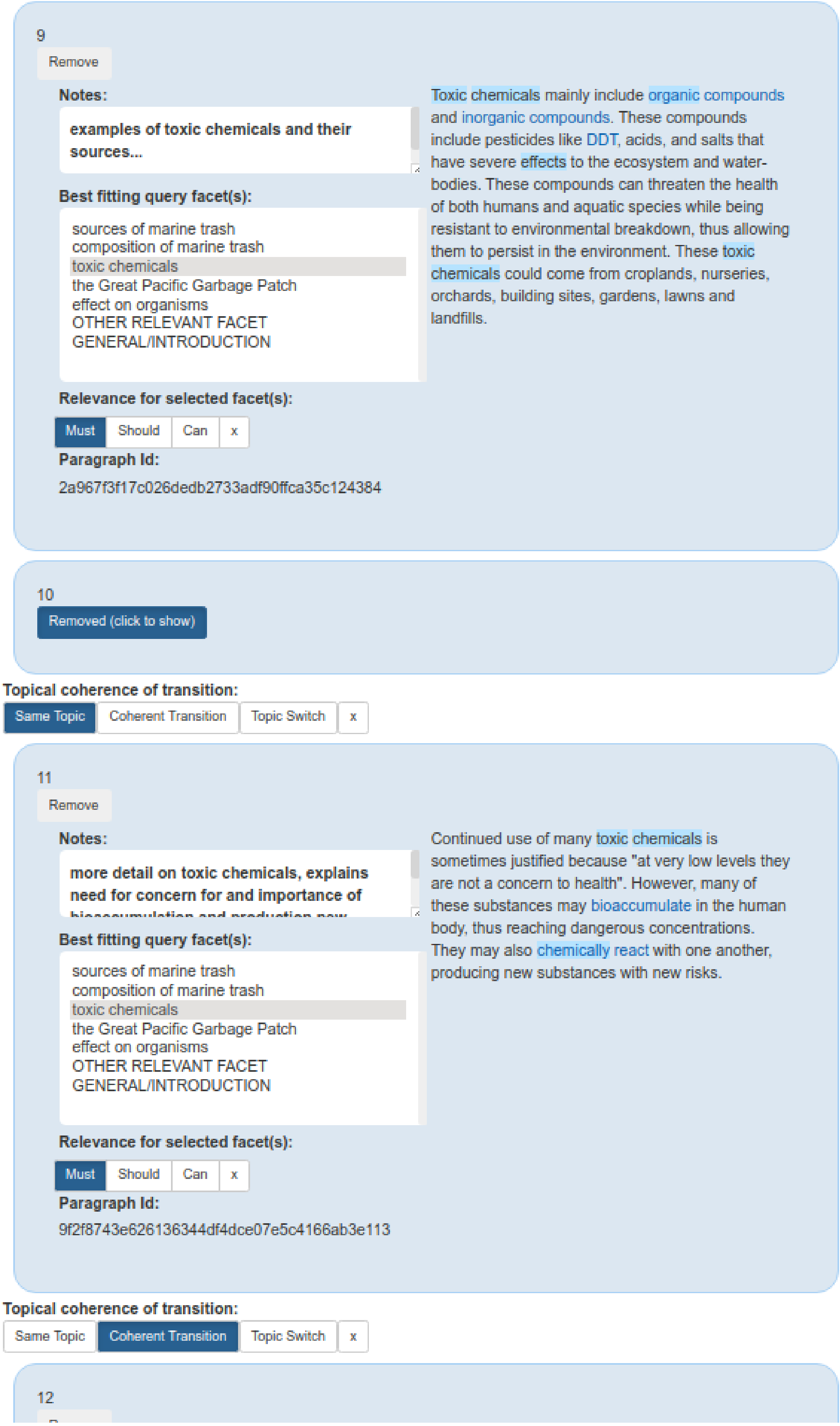
1. reading the passage
2. taking notes about if and why this passage is relevant for the query in the notes field.
3. detemining the best fitting facet for the passage and assigning a relevance label for how relevant this passage is for the selected facet, i.e., whether the passage MUST, SHOULD, or CAN be included on the generated article.
4. if the passage is not relevant for any facet, but should be displayed on the article (i.e., fitting the overarching theme), then judges were asked to select OTHER RELVANT FACET; if the passage a very relevant description of the topic, the judges were asked to select GENERAL/INTRODUCTION.
5. if the passage is not relevant for the article at all (i.e., the judge would have preferred to be shown this passage), the judges are asked to click the "Remove" button. – In this case the passage is hidden from the page view (In the screenshot, passage number 10 was removed)
6. For every pair of consecutive passages (skipping removed passages), the judged are asked to judge the smoothness of the transition. Possible choices are SAME TOPIC / COHERENT TRANSITION / TOPIC SWITCH. (In the screenshot the transition from passage number 9 to 11 was assessed, because 10 was removed.)

Detailed results of the assessments are provided in JSON format, see Appendix A.

5.2 Qrels

As the fast majority of submitted runs were directly derived from an ad hoc passage ranking, we also provide a qrels file for rank quality evaluation. We use the same scale as in previous years.

- 3: MUST be mentioned
- 2: SHOULD be mentioned
- 1: CAN be mentioned
- 0: Removed, non-relevant



9

Remove

Notes:

examples of toxic chemicals and their sources...

Best fitting query facet(s):

sources of marine trash
composition of marine trash
[✓ toxic chemicals]
the Great Pacific Garbage Patch
effect on organisms
OTHER RELEVANT FACET
GENERAL/INTRODUCTION

Relevance for selected facet(s):

[✓ Must] Should Can x

Paragraph Id:

2a967f3f17c026dedb2733adf90ffca35c124384

Toxic chemicals mainly include organic compounds and inorganic compounds. These compounds include pesticides like DDT, acids, and salts that have severe effects to the ecosystem and water-bodies. These compounds can threaten the health of both humans and aquatic species while being resistant to environmental breakdown, thus allowing them to persist in the environment. These toxic chemicals could come from croplands, nurseries, orchards, building sites, gardens, lawns and landfills.

10

Removed (click to show)

Topical coherence of transition:

[✓ Same Topic] Coherent Transition Topic Switch x

11

Remove

Notes:

more detail on toxic chemicals, explains need for concern for and importance of bioaccumulation and production new

Best fitting query facet(s):

sources of marine trash
composition of marine trash
[✓ toxic chemicals]
the Great Pacific Garbage Patch
effect on organisms
OTHER RELEVANT FACET
GENERAL/INTRODUCTION

Relevance for selected facet(s):

[✓ Must] Should Can x

Paragraph Id:

9f2f8743e626136344df4dce07e5c4166ab3e113

Continued use of many toxic chemicals is sometimes justified because "at very low levels they are not a concern to health". However, many of these substances may bioaccumulate in the human body, thus reaching dangerous concentrations. They may also chemically react with one another, producing new substances with new risks.

Topical coherence of transition:

Same Topic [✓ Coherent Transition] Topic Switch x

12

Figure 1: Screenshot of Assessment Interface.

6 Submission

Every team was allowed to submit up to 10 runs of different assessment priorities. Three runs from each team were fully assessed (priority HIGH and MEDIUM). Runs needed to be submitted in a JSON format that at the minimum needed to specify the paragraph ids in topically coherent order. The file was allowed to also include rankscore information for a more traditional ad hoc retrieval assessment.

We provided a script for converting Y2-style rankings into the new format, simply by taking the highest ranked passages in order of relevance for each section, then concatenating resulting paragraphs.

Unfortunately, all submitted runs (except a single submitted run) were created with this script. Hence a detailed evaluation of topical coherence would not make sense.

7 Results

Teams DANGNT-NLP, ECNU, IRIT, Smith, ICTNET, TREMA-UNH, and UAmsterdam participated this year. The results are presented in Figure 2. More detailed results will be available online.[1]

The three top ranked systems all make use of BERT and anserini, following Nogueira et al [8]. The next three methods are based on Terroer with a CombMNZ combination. Many remaining runs include different variation of BM25, RM3, and reranking. However, just using BERT is not a guarantee for good retrieval performance.

This is a lesson learned at last year's TREC meeting, where Nogueira's submission did not use an English stemmer when retrieving candidate pools with BM25, which severely impacted the performance of team NYU. This year, the community successfully employed Nogureira's BERT-based method [8] for the CAR passage ranking task.

Acknowledgement

We express our gratitude for many suggestions of several experts in the field, who helped to make this track successful. We thank the University of New Hampshire for providing computational resources and web servers. We are deeply thankful for Ellen Voorhees' experience, patience, and persistence in running the assessment process. Finally we thank all our participants.

References

[1] Jeffrey Dalton, Shahrzad Naseri, Laura Dietz, and James Allan. Local and global query expansion for hierarchical complex topics. In *European Conference on Information Retrieval*, pages 290–303. Springer, 2019.

[2] Jeffrey Dalton, Chenyan Xiong, and Jaime Callan. Cast 2019: The conversational assistance track overview. In *Text REtrieval Conference (TREC)*, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[4] Laura Dietz. ENT rank: Retrieving entities for topical information needs through entity-neighbor-text relations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 215–224. ACM, 2019.

[5] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[1] `http://trec-car.cs.unh.edu/results-Y3/`

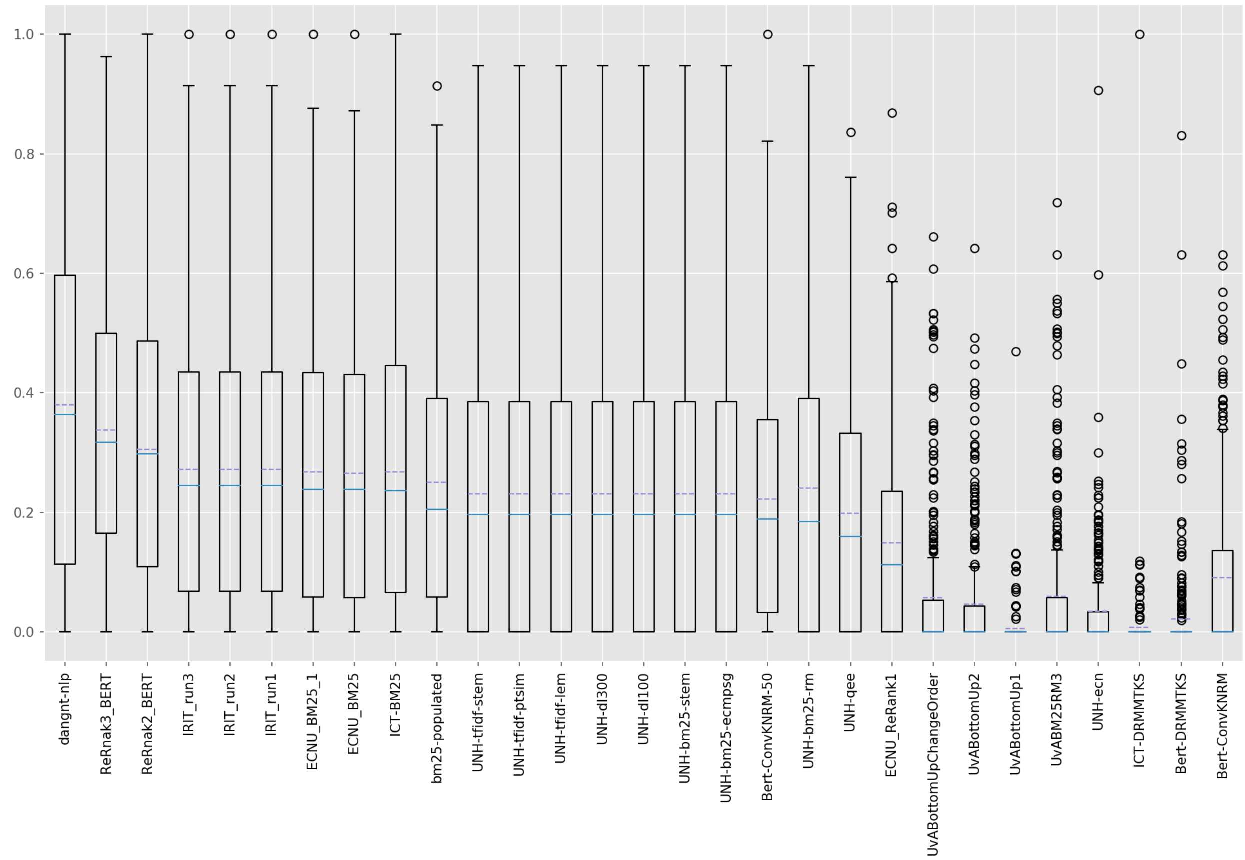


Figure 2: Section-level ranking performance in NDCG.

[6] Sean MacAvaney, Andrew Yates, Kai Hui, and Ophir Frieder. Content-based weak supervision for ad-hoc re-ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 993–996. ACM, 2019.

[7] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.

[8] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*, 2019.

A Format for Detailed Relevance Assessments

Relevance data is provided in JSON format describe below. There is a separate file for each query/judge. Each JSON file has the following entries (each representing a keyed on the pair `query_id, paragraph_id`):

- `notes_state` : notes for each paragraph from assessment step 2)
- `facet_state` : facet/relevance assessment from assessment step 3). Lists the most relevant `facet` with `heading` and `heading_id` (technically this is a section path in the form of title/heading) and the assessed `relevance` ranging over values MustLabel, ShouldLabel, CanLabel; special facet ids from step 4) are indicated by the `heading_ids`: `NONE_OF_THESE` and `GENERAL/INTRODUCTION`.
- `nonrelevant_state` : not used.
- `nonrelevant_state2` : contains `query_id`, `paragraph_id` that are marked as “REMOVE” in assessment step 5)
- `transition_label_state`: contains transition judgments from Step 6). Entries are keyed on `query_id, paragraph_id1, paragraph_id2`, and contain values `SameTransition`, `AppropriateTransition`, or `SwitchTransition`. — Paragraphs that were marked as “REMOVED” are skipped in transition assessments.

Each of these entries are represented as association maps from a pair of `query_id, paragraph_id` to a list of wrapped `value`. The wrapper also contains metadata of the assessment (`annotator_id, time_stamp, session_id, run_ids`). Since JSON does not support maps of complex keys, we represent each key-value pair in the map a list of length 2.

Example of `notes_state` and `facet_state` for passage number 9 (see Figure 1).

```
{
  "notes_state": [
    [
      {
        "query_id": "tqa2:L_0257",
        "paragraph_id": "2a967f3f17c026dedb2733adf90ffca35c124384"
      },
      [
        {
          "annotator_id": "NIST",
          "time_stamp": "2019-10-11T21:02:58.354696907Z",
          "session_id": "CAR-Y3",
          "run_ids": [],
          "value": "examples of toxic chemicals and their sources..."
        }
      ]
    ], ...
  ],
  "facet_state": [
    [
      {
        "query_id": "tqa2:L_0257",
        "paragraph_id": "2a967f3f17c026dedb2733adf90ffca35c124384"
      },
      [
        {
          "annotator_id": "NIST",
          "time_stamp": "2019-10-11T21:02:58.354696907Z",
          "session_id": "CAR-Y3",
          "run_ids": [],
          "value": {
            "facet": {
              "heading": "toxic chemicals",
              "heading_id": "tqa2:L_0257/T_1512"
            },
            "relevance": "MustLabel"
          }
        }
      ]
    ],...
   ],
```

Example of non-relevant_state2 for passage number 10

```
  "nonrelevant_state2": [
    [
      {
        "query_id": "tqa2:L_0257",
        "paragraph_id": "f1dec26869a4c7d00c9acee595d1dfa8afa69ffe"
      },
      {
        "annotator_id": "NIST",
        "time_stamp": "2019-10-11T21:02:58.354696907Z",
        "session_id": "CAR-Y3",
        "run_ids": [],
        "value": true
      }
    ], ...
  ],
```

Example of transition assessmend between passage number 9 and number 11

```
  "transition_label_state": [
    [
      {
        "query_id": "tqa2:L_0257",
        "paragraph_id1": "2a967f3f17c026dedb2733adf90ffca35c124384",
        "paragraph_id2": "9f2f8743e626136344df4dce07e5c4166ab3e113"
      },
      {
        "annotator_id": "NIST",
        "time_stamp": "2019-10-11T21:02:58.354696907Z",
        "session_id": "CAR-Y3",
        "run_ids": [],
        "value": "SameTransition"
      }
    ], ...
  ]
}
```