TREC 2024 NeuCLIR Track Guidelines
Pilot Report Generation Task

Version 1.0, March 24, 2024
Updated: May 28, 2024
For CLIR and MLIR tasks, see CLIR and MLIR Task Guidelines.
Further information about the TREC 2024 NeuCLIR Track can be found on the track Website.
Please leave your feedback on these guidelines as comments in this document.

TL;DR

Participating systems will receive a Chinese, Persian, or Russian document collection and a request for a report to be written in English. They will produce a textual report that fulfills the request in which each sentence points to up to two documents that support it. Scoring will be based on the percentage of main points the report successfully includes, and the precision of the report sentences in describing those main points.

Task

NeuCLIR 2024 is introducing a new Retrieval Augmented Generation (RAG) task of producing a report in one language based on the retrieval of documents in another language. The goals of the task are to stimulate research in cross-language RAG, to identify the best current approaches to automatic cross-language report-writing, to produce a collection that is useful for evaluating future systems, and to examine what parts of this evaluation might be automated to allow future systems to be scored using the same evaluation data. This year report generation is a pilot task, so participants might expect some details of the task to change slightly over time.

The report generation task is to automatically generate a report based on an English description of the desired report and a document collection in Chinese, Persian, or Russian. Generated reports must be responsive to the statements of the information need. A valid report will cite source documents that contain the information discussed within the report. Citations are one of the key attributes of this task. A report length limit will encourage systems to express information succinctly. The generated report must be written in the language of the report request (English) rather than in the language of the documents. Submitted reports created by report-writing systems will be evaluated by assessors.

Inputs

The first system input is the document collection that will be used as source material for the retrieval task. Following most information retrieval-based evaluations, documents are assumed to be truthful; verifying the truthfulness of document contents is beyond the scope of the evaluation.

The second system input is a set of information needs, which are developed by assessors and are referred to as *report requests.* A report must be generated for each report request. For the pilot report requests will include:

- Background: The report requester's background.
- Problem statement: A description of the content the report is required to contain.
- Length restriction: Maximum number of Unicode characters the report may include after NFKC normalization. Report sentences that contain characters beyond this limit will be truncated before the report is scored. The `normalize` method in the `unicodedata` Python package will perform this normalization.

The report request will be expressed in unstructured text except for the length constraint. Here is an example report request:

> **Background**: I am a Hollywood reporter writing an article about the highest grossing films Avengers: Endgame and Avatar.
>
> **Problem statement**: The article needs to include when each of these films was considered the highest grossing film and any manipulations undertaken to bring moviegoers back to the box office with the specific goal of increasing the money made on the film.
>
> **Limit**: 2000

For the pilot, we intend the report request topics to align with the CLIR topics, and will use the same topic IDs for the two tasks. Whether it proves possible to maintain the same definition of document relevance across those tasks and whether the relationship is useful either for report generation or CLIR remains to be seen; we welcome comments on this relationship.

Output

The report will be generated by a *report writer* that is an automated system. Reports produced by the report writer must satisfy all aspects of the report request.

Each generated report must be segmented into sentences, either manually or automatically, by the submitting team. The track will provide code to segment sentences and formulate the result in the required format (with empty citation lists); however, a participating team is

welcome to use their own sentence segmentation. If necessary, participants may correct any segmentation errors in a run without the run being counted as a manual run (bearing in mind that no changes to the system software may be made after doing so).

The report must also include appropriate *citations*, pointers from individual sentences to supporting documents, as described below

Citations

Each substantive sentence of a submitted report must cite the document(s) from which it can be derived. A citation is a document ID paired with a report sentence. A given report sentence may bear more than one citation with a **maximum of two for the pilot**, and a given document may be cited more than once.

The validity of a citation has three components.

- First, the documents cited by a sentence should together support the sentence. That is, reading the cited documents should verify the sentence's accuracy.
- Second, the sentence bearing the citation should be responsive to the report request. The methodology for deciding whether a sentence is responsive to the report request is based on *nuggets,* which are described below.
- Third, the pilot evaluation will include an assessment of citation appropriateness: whether a talented author in the field of the report would include that citation had they written the report. The assessor will identify sentences that do not require citations. Cases where no citation is required include introductory sentences, background sentences that reflect the problem statement, conclusions drawn from other cited sentences, and summary sentences. If a citation is not needed and no citation is present, the report will be scored as if the identified sentence was not present in the report (that is, such cases will not affect the score).

Data Formats

Input Format (Report Requests)

The document collections for the Report Generation task will be the same NeuCLIR1 collections used for the other NeuCLIR tasks. See the CLIR/MLIR guidelines for the document format. If you would like to participate in the task but do not have access to a CLIR system, see the TREC webpage for how to access a retrieval service.

Report requests will be distributed in JSONL format as a list of individual requests one per line. Each request will contain the following JSON fields:

- `request_id` (string): A unique ID for this report request
- `collection_ids` (array): list of collection IDs to be searched. For the pilot, there will always be a single entry, either `neuclir/1/fas`, `neuclir/1/rus`, or `neuclir/1/zho`.

- `background` (string): Describes the context in which the report is being written
- `problem_statement` (string): Describes what should and should not be included in the report
- `limit` (int): Maximum number of NFKC-normalized Unicode characters the report may include

Output Format (Reports)

The submission format is a sequence of JSONL entries each representing one report. Each report is a JSON object containing:

- `request_id` (string): The unique ID of the input report request
- run_id (string): An arbitrary string to identify the run. It is recommended to include your team name as part of the run_id
- `collection_ids` (array): list of collection IDs that were searched. Should match the report request and be exactly one of `neuclir/1/zho`, `neuclir/1/rus`, or `neuclir/1/fas`.
- `sentences` (array): a list of sentence structures

Sentences must appear in report order. Each sentence structure has the following fields:

- `text` (string): a string containing the text of the sentence
- `citations` (array): a list of zero to two document IDs (strings)

Together, the `text` fields of each report must not exceed 2000 NFKC-normalized Unicode characters. An example report in the JSON format[1] follows:

```
{
      "request_id": "300",
      "run_id": "zho-myTeam-ANDG-anAwesomeReportGenerator",
      "collection_ids": ["neuclir/1/zho"],
      "sentences":[
            {"text": "Avengers: Endgame and Avatar are two of the
      highest-grossing films in history.", "citations": []},
            {"text": "Avengers: Endgame surpassed Avatar as the
      highest-grossing film globally, with a box office revenue of
      $2.787 billion.", "citations": ["D12"]}
            ⋮
            {"text": "These manipulations aimed to increase the films'
      box office revenue and solidify their positions as
      record-breaking blockbusters. ", "citations": []}
      ]
}
```

[1] This would appear as a single line of JSONL in an actual submission.

Data

See the track's Webpage for sample development data. The development data includes both report requests in the format described above as well as nugget questions with answers (see Appendix). There is currently no automated software for using the sample evaluation data to score reports.

Software

Please see CLIR and MLIR Task Guidelines for other relevant software.

Retrieval API

We will provide a retrieval API for participants who want to work on the report generation tasks but do not have the capacity to set up their own search engine. Service bandwidth is limited, so the service is not public. Please see the track description on the TREC website once you register for API details.

Document Length Tool

The Python script `reportlength.py`, available on the track Website, will report the length of each report in a JSONL submission file.

Sentence Segmentation

Participants with existing RAG systems that need to divide their system output into sentences may use any of the freely available packages, such as spaCy, NLTK, or CoreNLP.

Submission

Runs will be submitted through the NIST submission system. Runs that do not pass validation will be rejected outright. Submitted runs will be asked to specify (a) task; (b) document collection; and (c) automatic (no human intervention beyond sentence segmentation) or manual (human intervention at any point).

Teams may make an arbitrary number of submissions. Each team must indicate the order in which they would like their submissions assessed. There is no guarantee that runs beyond the first three of these will be included in pools or scored. Run IDs are an arbitrary string to identify the run. It is recommended to include your team name as part of the run_id.

Assessment & Scoring

Report assessment for each topic will be made by a single person per language. Cited documents will be added to the relevance assessment pools for the analogous CLIR topic when adding documents to the pools, assuming such a topic exists. After retrieval relevance assessment of pools, but before report assessment, assessors will create the required assessment data based on nugget questions and answers linked to the document collection. More details appear in the Appendix.

At least two measures will be used to score reports: recall over nuggets and precision over sentences. Nugget "question recall" is the fraction of nugget questions that have a valid answer. Sentence "citation precision" is the fraction of substantive sentences supported by cited content. Multiple citations can cover different facts in a single sentence. The process for determining the values of these measures is outlined in the Appendix.

Registration

Register starting March 1, 2024 at http://ir.nist.gov/evalbase.
NIST has posted the full call for participation for all 2024 Tracks at TREC.

Important Dates

March 2022:	Evaluation document collection released
Mar 25, 2024:	Track guidelines released
June 2024:	Report requests released
July 28, 2024:	Submissions due to NIST
October 2024:	Results distributed to participants
November 2024:	TREC 2024

Coordinators

Dawn Lawrie, Johns Hopkins University, HLTCOE
Sean MacAvaney, University of Glasgow
James Mayfield, Johns Hopkins University, HLTCOE
Paul McNamee, Johns Hopkins University, HLTCOE
Douglas W. Oard, University of Maryland
Luca Soldaini, Allen Institute for AI
Eugene Yang, Johns Hopkins University, HLTCOE

Appendix: Evaluation

The evaluation will focus on whether and how well the sentences of the report support a set of *nuggets* that encapsulate the information that a good report should provide. This appendix is presented only to help participants understand the evaluation; it does not represent guidelines to be followed.

Nuggets

A nugget is a piece of information identified by the assessor that should appear in the report, and that can be expressed in a variety of ways in the document collection. More formally, a nugget is a combination of a question that addresses some aspect of the report request and one or more answers to that question that are expressed in at least one document in the collection and that are deemed to be acceptable answers by the assessor. Nuggets will be expressed by the assessors at an appropriate level of granularity for the desired report. If the report answers a nugget question using appropriate citations into the document collection, we deem it to have succeeded in identifying that nugget. Answers to questions will express the information that a reasonable person would expect in a report written in response to the report request. Note that the nugget questions and answers are not available to participating systems; they are identified by the assessor and serve only to facilitate evaluation.

Nugget Identification

The concept of nuggets is not new and arose from summarization evaluation; what is new in this task is the expression of nuggets as questions with allowable answers. Nuggets need not capture everything any report responding to the report request might legitimately include. The assessor will determine the required information and express that information as nuggets, reinforcing the idea that nuggets, like relevance judgments, are opinions rather than facts. The pyramid method will not be used. The set of answers to a nugget question are drawn from all the answers supported by the document collection. Questions and answers will be expressed in the report request language (English), not the language of the documents.

Nuggets are identified by the assessor and are not revealed to participants during the evaluation. Nuggets must be both relevant to the report request and attested in the document collection. In addition to identifying the set of nuggets for a report request, the assessor will use their own searches, pools from the corresponding retrieval task, and documents cited by any submitted reports to identify nugget answers.

Any answer to a nugget question expressed in a document links that document to the corresponding nugget. Multiple answers may be possible if different documents express essentially the same information in different ways (consider “2021” vs. “two years after the release of Avengers Endgame”). During report assessment, annotators may add additional nugget answers, but no new nugget questions will be added. Nuggets with multiple answers that all should be present in a report will be counted as multiple individual nuggets for

scoring purposes. A report sentence will be evaluated based on whether it provides an answer to a nugget question that is acceptable to the assessor using a known citation. The assessor will determine whether the report sentence answers one of the nugget questions.

Scoring Sentences

Figure 1 expresses the pilot approach to sentence scoring. In the flowchart, a "+" means the sentence is correct and should be rewarded; a "–" means that it is incorrect and should be penalized; and a "0" means that the sentence is not included when scoring the document. The flowchart shows how each sentence of the report can be scored. We will use the following principles to guide sentence scoring:

- Sentences with citations whose cited document does not support them will be penalized (Score #1 in Figure 1).
- Properly cited and attested sentences that are not relevant to the report will be ignored (Score #2).
- A sentence that cites a document supporting a nugget that the sentence fulfills will be rewarded (Score #3).
- Sentences that neither have nor require citations will not affect the score (Score #4).
- Sentences that should contain a citation will only be penalized once for not citing a particular fact (i.e., a repeated fact that should have citation will only be penalized once) (Scores #5 and #6).
- Sentences that claim the absence of a fact will be rewarded or penalized depending on whether the absence is explicitly stated as a nugget (Scores #7 and #8). A nugget will be created for any information that the report request explicitly asks for but is not attested in the collection.

Most sentences will have either zero or one citation. It is allowable for a sentence to have multiple citations (up to two in this pilot), either because the same information is attested multiple times in the collection, or because it is a complex sentence that requires multiple citations. Sentences that cite multiple documents that support the same nugget are simply treated as a single citation. The evaluation will macroaverage the citation scores to ensure all sentences are given equal weight. From the nugget perspective, support by multiple report sentences is counted only once.

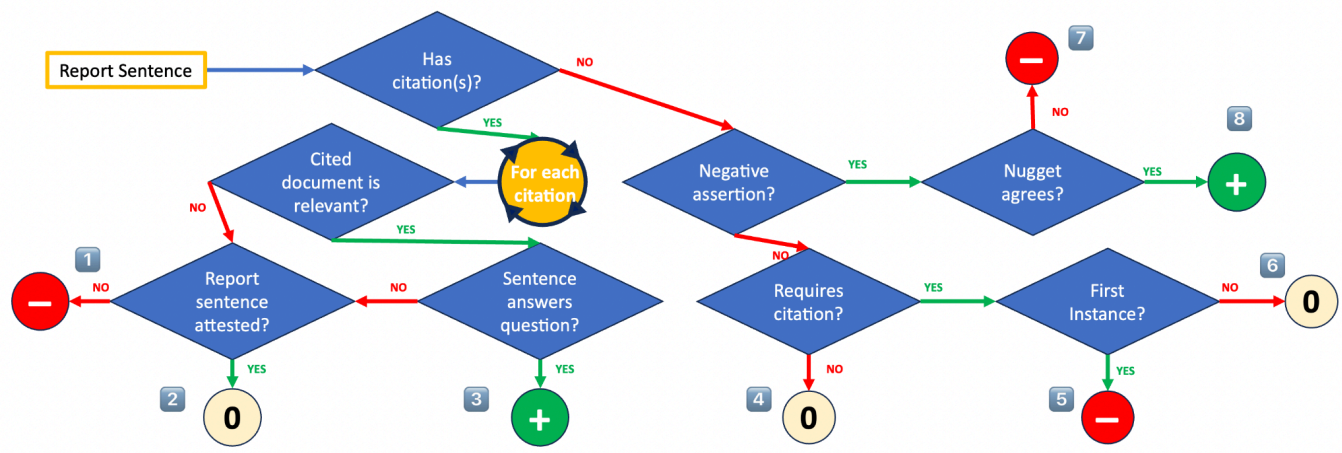

Figure 1: Flowchart for sentence evaluation

Metrics

The report evaluation pilot will use nugget recall and sentence precision. We focus on these two measures here because they are well-known, easy to calculate, and highlight most of the important scoring issues we face in generated report evaluation.

To calculate nugget recall and sentence precision, we need a numerator and a denominator for each. The nugget recall denominator is the number of distinct assessor-identified nuggets. The numerator is the number of correctly reported nuggets. A nugget is correctly reported if the report contains one or more well-formed citations to a document that answers a nugget question (that is, matches one or more of the nugget canonical answers). So the recall measure tells us how many of the concepts central to the report were actually reported on.

The sentence precision denominator is the number of report sentences, minus any sentence that does not require a citation or that properly cites information not identified in the nuggets. The numerator is the number of sentences deemed to bear accurate citations.

Example Assessment

Figure 2 shows an example of the two types of information needed to do manual or automatic assessment. The report request is the input that will be used by the report writer to create the report. The nugget questions along with acceptable answers show how each answer is linked to the documents that attest to that answer.

Figure 3 is a report generated in response to the example in Figure 2, broken into report sentences to illustrate manual evaluation. Each Score: # indicates how the sentence would be categorized using the flowchart in Figure 1. For Score: #3, the nugget answer in the sentence is also recorded.

There are a few points to make about this particular generated report. First, no sentence received a negative assessment because there were no scores of #1 or #7. Therefore, its precision score is 5/(16 − 11) = 1.0. For recall, one nugget was repeated three times in Lines 10, 11, and 12, so its recall score is 3/5 = 0.6. Finally, for both Line 2 and Line 6, the assessor would need to refer to the original source document to assess the statement, since the information appearing in the sentence had not been captured in a required nugget.

Report Request:

Background: I am a Hollywood reporter writing an article about the highest grossing films Avengers: Endgame and Avatar.

Problem Statement: The article needs to include when each of these films was considered the highest grossing films and any manipulations undertaken to bring moviegoers back to the box office with the specific goal of increasing the money made on the film.

Limit: 2000

Sample Report: Avatar originally became the highest grossing film in 2010 [D1]. Avengers: Endgame replaced Avatar as the highest grossing film in 2019 [D1, D10]. It overtook Avatar by adding an additional six minutes of footage to the film to draw viewers back to the movie theater [D4]. Two years later Avatar was re-released in mainland China [D1, D11]. It earned a sufficient amount of money to retake the title of highest-grossing film in 2021 [D5, D11].

Nugget Questions and Answers:

1. When did Avatar first become the highest grossing film?
    - 2010 [D1]
2. When did Avengers: Endgame become the highest grossing film?
    - 2019 [D1,D2, D3, D8, D10, D12, D13]
    - July 2019 [D3, D12, D13]
    - July 20, 2019 [D3]
    - July 21, 2019 [D13]†
3. What did studio executives do to the Avengers: Endgame film to become the highest grossing film?
    - Added six minutes of additional footage [D4]
    - Added footage [D4]
    - Added 6 minutes [D4]
    - Additional footage at the end of the film [D14]
4. When did Avatar retake the title of highest grossing film?
    - 2021 [D1, D2, D6,D7,D9,D11]
    - March 2021 [D1, D6 ,D7, D9, D11]
    - March 13, 2021 [D1, D6, D9]
    - Two years after the Avengers: Endgame became the highest grossing film [D2]
5. What event led to Avatar becoming the highest grossing film?
    - Re-release in Mainland China [D1, D2, D5, D6, D7, D8, D9, D10]
    - Re-release in China [D1, D2, D5, D6, D7, D8, D9, D10]
    - Release in Mainland China for a second time [D1, D2, D5, D6, D7, D8, D9, D10]Returned to theaters in China [D11]

†In Taiwan Time

Figure 2: Example evaluation material for a report request

1. Avengers: Endgame and Avatar are two of the highest-grossing films in history. `Score: #4`
2. Avengers: Endgame surpassed Avatar as the highest-grossing film globally, with a box office revenue of $2.787 billion. [D12] `Score: #2`
3. This achievement was officially announced in July 2019. [D13] `Score: #3, Nugget: 2`
4. To bring moviegoers back to the box office and increase the revenue of Avengers:Endgame, Marvel Studios implemented several strategies. `Score: #4`
5. One of these strategies was re-releasing the film with additional content. `Score: #6`
6. Marvel Studios President Kevin Feige confirmed plans for a re-release, including previously unseen footage, to generate renewed interest and boost ticket sales. [D101] `Score: #2`
7. The new version of the film did not include any re-edited scenes within the main feature but instead featured additional scenes after the end credits, including a tribute. [D14] `Score: #3, Nugget: 3`
8. This approach aimed to entice both existing and new audiences to revisit the theaters and contribute to the film's box office success. `Score: #6`
9. Similarly, Avatar also made a comeback to reclaim its position as the highest-grossing film. `Score: #4`
10. The film was re-released in China, where it achieved a single-day box office revenue of $8 million. [D2] `Score: #3, Nugget: 5`
11. The re-release in China was particularly successful, as it generated significant revenue and propelled Avatar back to the top of the global box office charts. [D1] `Score: #3, Nugget: 5`
12. The film's re-release in China was attributed to the impact of the COVID-19 pandemic, which led to the delay of many new releases and created an opportunity for Avatar to regain its popularity. [D2] `Score: #3, Nugget: 5`
13. In conclusion, Avengers: Endgame and Avatar both held the title of the highest-grossing film globally at different points in time. `Score: #4`
14. Marvel Studios strategically re-released Avengers: Endgame with additional content to attract audiences and boost ticket sales. `Score: #6`
15. Avatar capitalized on the re-release trend in China, taking advantage of the pandemic-induced delay of new releases. `Score: #6`
16. These manipulations aimed to increase the films' box office revenue and solidify their positions as record-breaking blockbusters. `Score: #4`

Figure 3: Example of report evaluation