TREC 2020 Misinformation - Evaluation Guidelines

August 18, 2020

In the following we provide a detailed description on how to compute evaluation measures for the Ad Hoc task of TREC 2020 Misinformation Track.

Qrels: The final qrels will contain assessments with respect to the following criteria:

- *Usefulness*: does this document contain material that the search user might find useful in answering the question? Usefulness will be assessed on a binary scale: 0 if the document is not useful in answering the question and 1 otherwise;
- *Answer*: does the document answer to the question in the description field? If so, is the answer yes or no? The answer will be assessed with 3 values: 0 means that the document does not answer the question, -1 means that the answer provided by the document is "no", and 1 means that the answer is "yes";
- *Credibility*: how credible is the document? Credibility will be assessed on a binary scale, where 0 stands for not credible and 1 stands for credible.

Note that non-useful documents will not be assessed with respect to credibility and wether they provide an answer.

The format adopted for the qrels file is as follows:

```
topic_id 0 doc_id usefulness-judgment answer-judgment credibility-judgment
```

where the columns are space separated, and the last three columns report usefulness, answer, and credibility labels respectively.

Mapping to Correctness: Before applying any evaluation measure, answer labels should be mapped to *correctness* labels. Correctness labels will be obtained with the topic answer field and the answer labels provided by assessors. In practice, given a topic with answer field equal to "yes", all documents assessed with answer 1 will be considered correct for that topic. Similarly, given a topic with answer field equal to "no", all documents assessed as -1 will be considered correct. Thus, the final qrels used for evaluation will be as follows:

```
topic_id 0 doc_id usefulness-judgment correctness-judgment credibility-judgment
```

where the answer label is replaced by correctness. Correctness will be encoded with a binary scale, where 0 stands for not correct and 1 stands for correct.

Single aspect Evaluation: We will evaluate each run by considering each aspect separately. In practice, we will consider in the qrels just the column corresponding to usefulness, correctness and credibility and use `trec_eval`[1] to compute Average Precision (AP) and Normalized Discounted Cumulated Gain (nDCG) with cut-off 10. We will refer to these measures as $\mathcal{M}_{use}$, $\mathcal{M}_{corr}$ and $\mathcal{M}_{cre}$ for usefulness, correctness and credibility, where $\mathcal{M}$ can be AP or nDCG@10.

[1] `https://github.com/usnistgov/trec_eval`

Multi-aspect Evaluation - Aggregation of Measures: To evaluate the runs with respect to the 3 aspects we will use Convex Aggregating Measure (CAM) [1] and Multidimensional Measure (MM) [2], instantiated with AP and nDCG@10.

Let r_t be a ranked list of documents with multi-aspect labels for a topic t, CAM is the convex sum of the $\mathcal{M}$ scores computed with respect to each aspect individually:

$$\text{CAM}(r_t) = \lambda_{use}\mathcal{M}_{use}(r_t) + \lambda_{corr}\mathcal{M}_{cor}(r_t) + \lambda_{cre}\mathcal{M}_{cre}(r_t) \tag{1}$$

where $\mathcal{M}_{use}$, $\mathcal{M}_{cor}$, and $\mathcal{M}_{cre}$ denotes respectively AP or nDCG@10 for usefulness, correctness, and credibility, and $\lambda_{use}+\lambda_{corr}+\lambda_{cre} = 1$ are non negative parameters controlling the impact of the individual usefulness, correctness and credibility scores in the overall computation. We will assume $\lambda_{use} = \lambda_{corr} = \lambda_{cre} = 1/3$.

Similarly, MM is the harmonic mean of of the $\mathcal{M}$ scores computed with respect to each aspect individually:

$$\text{MM}(r_t) = \frac{1}{\frac{\lambda_{use}}{\mathcal{M}_{use}(r_t)} + \frac{\lambda_{cor}}{\mathcal{M}_{cor}(r_t)} + \frac{\lambda_{cre}}{\mathcal{M}_{cre}(r_t)}} \tag{2}$$

where we will assume $\lambda_{use} = \lambda_{corr} = \lambda_{cre} = 1/3$.

Multi-aspect Evaluation - Aggregation of Labels: Finally, we will aggregate labels first and then compute AP and nDCG@10, opposed to CAM and MM which compute measures scores and then average scores across aspects. To aggregate labels we will use 2 different strategies:

- *Harsh:* we will consider the minimum label across aspects:

  ```
  topic_id 0 doc_id min{usefulness-judgment,correctness-judgment,credibility-judgment}
  ```

  For example a document which is labelled `1 0 1` will have an aggregated label of 0. Therefore, only a document which is useful, correct and credible will contribute to the measure score;

- *Lenient:* we will sum the labels across aspects:

  ```
  topic_id 0 doc_id sum{usefulness-judgment,correctness-judgment,credibility-judgment}
  ```

  For example a document which is labelled `1 0 1` will have an aggregated label of 2. Therefore, documents contribute to the measure scores depending on how many criteria they fulfil.

Given the new qrels with the aggregated labels, we will use `trec_eval` to compute AP and nDCG@10 for both the aggregation strategies.

References

[1] C. Lioma, J. G. Simonsen, and B. Larsen. Evaluation Measures for Relevance and Credibility in Ranked Lists. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR 2017, pages 91–98, New York, NY, USA, 2017. ACM.

[2] J. Palotti, G. Zuccon, and A. Hanbury. MM: A new Framework for Multidimensional Evaluation of Search Engines. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM 2018, pages 1699–1702, New York, NY, USA, 2018. ACM.