# Overview of the TREC 2019 Decision Track

Mustafa Abualsaud[1], Christina Lioma[2], Maria Maistro[2], Mark D. Smucker[1], and Guido Zuccon[3]

[1]University of Waterloo
[2]University of Copenhagen
[3]University of Queensland

## 1 Introduction

Search engine results underpin many consequential decision making tasks. Examples include people using search technologies to seek health advice online [10, 18], or time-pressured clinicians relying on search results to decide upon the best treatment/diagnosis/test for a patient [22, 20].

A key problem when using search engines in order to complete such decision making tasks, is whether users are able to discern authoritative from unreliable information and correct from incorrect information. This problem is further exacerbated when the search occurs within uncontrolled data collections, such as the web, where information can be unreliable, generally misleading, too technical, and can lead to unfounded escalations [24]. Information from search engine results can significantly influence decisions, and research shows that increasing the amount of incorrect information about a topic presented in a Search Engine Result Page (SERP) can impel users to take incorrect decisions [16]. As noted in the SWIRL III report [7], decision making with search engines is poorly understood, and likewise, evaluation measures for these search tasks need to be developed and improved.

In this context, the TREC 2019 Decision track aims to (1) foster research on retrieval methods that promote better decision making with search engines, and (2) develop new online and offline evaluation methods to predict the decision making quality induced by search results.

This overview paper is organized as follows: Section 2 describes the track setup, the collection and the official evaluation measures, Section 3 reports and discusses the evaluation results for the submitted runs, and finally Section 4 outlines future directions for the next edition of the track.

## 2 Decision Track Setup

The track is planned over multiple years, with data and resources created in one year flowing into the next year. We plan for the track to run for at least 3 years, with 2019 being the first year.

In this year's edition, we proposed only Task 1, where we asked participants to devise search technologies that promote correct information over incorrect information, with the assumption that correct information can better lead people to make correct decisions. Note that this task is more than simply a new definition of what is relevant. Because incorrect information can have a negative effect on decisions [16], there are three types of results: correct and relevant, incorrect, and non-relevant. It is important that search results avoid incorrect results, and ranking non-relevant results above incorrect is preferred.

For the next year's edition of the track (2020), we are renaming the track to be the Health Misinformation Track. The goals of the track remain the same. The new name makes clear that the key problem we are focused on in the near term is that of how health misinformation in search can negatively affect searcher decisions. For the TREC 2020 Health Misinformation track, we will add two tasks in addition to the current retrieval task. As planned, a new task will be to predict user decisions after the search in an offline context, i.e. to develop new evaluation measures for the track. The existing and continuing retrieval task (Task 1)

has a goal of finding relevant, credible, and correct information, and to complement this task, we will add a task of finding all of the documents containing misinformation for a given search topic. Further details about the next year edition can be found in Section 4.

In the following sections, we describe the corpus and topics used in Task 1, and the assessment phase performed by National Institute of Standards and Technology (NIST).

## 2.1 Corpus and Topics

**Corpus** The track used ClueWeb12-B13 [1] as the corpus, since this web collection provides a readily available source of documents that contain both correct and incorrect information and documents of varying credibility and quality. The full dataset consists of 52,343,021 English Web pages, collected between February 10, 2012 and May 10, 2012, and is a representative 7% sample of the whole ClueWeb12 corpus.

**Topics** The track focused on topics within the consumer health search domain (people seeking health advice online) to form user stories (search topics). Consumer health search represents an ideal prototypical example of the consequential decisions that we want search engines to correctly support. This domain also allows us to use evidence from systematic reviews [2] [8] to inform what a correct decision may be. Previous information retrieval evaluation challenges have tackled problems related to consumer health search, e.g., the CLEF eHealth CHS tasks [14, 12] and the FIRE 2016 CHIS task [17]. However the TREC Decision Track is novel in that it goes beyond the retrieval and ranking of search results, but it also consider the consequent decisions people make based on this information. This TREC track also considers multiple aspects of relevance along with topicality, and specifically correctness and credibility – in this it is similar to the CLEF eHealth CHS track, that considered trustworthiness (which partially resembles the notion of credibility we consider here) and understandability.

Mark Smucker's research group at the University of Waterloo developed the set of topics for the track. Each topic consisted of a health treatment and a health issue (e.g. acupuncture for insomnia). Some of the topics used in the track are the same as used by White and Hassan [23]. White and Hassan [23] categorized the topics into three categories depending on the effectiveness of the treatment towards the health issue. The three categories are summarized as follows:

- *Helpful*: The health treatment is helpful towards the health issue.
- *Inconclusive*: It is still unclear by medical professionals whether or not the treatment is effective towards the health issue.
- *Not helpful*: The treatment is not helpful towards the health issue.

White and Hassan [23] assessed the effectiveness of the health treatments of their topics by reading the corresponding Cochrane Review [3]. Cochrane Review provides systematic reviews of health-care interventions made by medical professionals for a broad audience [3]. Other topics were selected from the Cochrane Review library, and were assessed in a similar way as in White and Hassan [23]. In total, we created 51 topics (17 for each of the three categories above). Examples topics are reported in Figure 1.

## 2.2 Relevance, Efficacy and Credibility Assessment

The documents were judged by NIST with respect to three aspects: topical relevance, efficacy and credibility. A total of 22 859 assessments were collected. Table 1 reports the labels that were collected for each aspect with the corresponding value in the qrels file and the percentage of documents for each label.

Assessors judged a document for efficacy and credibility only if that document has been judged as Highly Relevant or Relevant.

The format adopted for the qrels file is as follows:

[1] https://lemurproject.org/clueweb12/

[2] Systematic reviews summarise the available scientific evidence towards an hypothesis, e.g. the effectiveness of a treatment for a specific condition.

[3] https://www.cochranelibrary.com/

```
<topic>
  <number>7</number>
  <query>aspirin vascular dementia</query>
  <cochranedoi>10.1002/14651858.CD001296</cochranedoi>
  <description>
    Can aspirin improve the lives of people with vascular dementia?
  </description>
  <narrative>
    Vascular dementia is a brain disorder that occurs as a result of dysfunction in the
    vascular system that carries blood to the brain. It is suggested that aspirin can help
    to improve the vascular system and benefit people with dementia. Relevant documents
    should discuss whether aspirin could be used as a treatment to help people with vascular
    dementia and reduce severity of its symptoms. Documents that don't discuss the
    effectiveness of aspirin for treating vascular dementia but discuss other dementia
    related issues such as Alzheimer and Lewy Bodies should be regarded as irrelevant.
  </narrative>
</topic>
<topic>
  <number>8</number>
  <query>melatonin jet lag</query>
  <cochranedoi>10.1002/14651858.CD001520</cochranedoi>
  <description>Can melatonin be used to reduce jet lag?</description>
  <narrative>
    Jet lag is a fatigue and sleep disorder caused by air travel across several time zones.
    It has been suggested that melatonin can be used to reduce or prevent the effects of jet
    lag. Relevant documents should discuss whether taking melatonin can be effective for
    treating jet lag.
  </narrative>
</topic>
<topic>
  <number>9</number>
  <query>ear drops remove ear wax</query>
  <cochranedoi>10.1002/14651858.CD012171.pub2</cochranedoi>
  <description>Can ear drops remove ear wax?</description>
  <narrative>
    Build up of ear wax can cause problems, e.g. hearing loss, and may require interventions
    such as syringing. Different types of ear drops have been suggested to be useful to
    soften ear wax and be used to remove it. A relevant document should discuss the
    effectiveness of any type of ear drops in removing ear wax.
  </narrative>
</topic>
```

Figure 1: Example topics.

```
topic_id 0 doc_id relevance-judgment efficacy-judgment credibility-judgment
```

where the columns are space separated, and the last three columns report the relevance, efficacy, and credibility labels respectively. Figure 2a shows a sample of the original qrels produced by NIST.

Note that relevance assessments were not collected for topic 14 due to time limitations from NIST. Topic 14 has therefore been excluded from the qrels and the submitted runs, and a total of 50 topics (rather than the original 51 provided to participants) is used for evaluation in the TREC 2019 Decision track.

**Topical Relevance Assessment** Topical Relevance was judged similar to previous tracks at TREC. However, unlike previous tracks, the assessors did not create their own topic statements; instead, they were provided the topic query and narrative as shown in Figure 1.

Based on [21], an assessor could judge the topical relevance of documents on a three points scale:

- *Highly Relevant*: if the document directly addresses the core issue of the topic;
- *Relevant*: if the document contains information that the user would find helpful in meeting their information need;
- *Not Relevant*: if the document does not contain helpful information, written in a foreign language (not in English), contains adult material, is unreadable or broken.

Note that the assessors were asked to judge the documents considering solely topical relevance, without considering whether the information provided by a document was incorrect and could harm the user.

Table 1: Labels collected for each aspect together with the corresponding value in the NIST qrels file and the percentage of documents with that value in the qrels. For "Efficacy – Not Judged" and "Credibility – Not Judged", the difference between −1 and −2 is as follows: −1 is assigned whenever the document was not relevant, while −2 is assigned whenever the document was relevant, but it was erroneously missed in the judged process by NIST.

| | Label | qrels value | % Judgments |
|---|---|---|---|
| Topical Relevance | Highly Relevant | 2 | 4.50 % |
| | Relevant | 1 | 13.72 % |
| | Not relevant | 0 | 81.78 % |
| Efficacy | Effective | 3 | 13.23 % |
| | Inconclusive | 2 | 3.85 % |
| | Ineffective | 1 | 0.70 % |
| | No Information | 0 | 0.38 % |
| | Not Judged | −1 or −2 | 81.84 % |
| Credibility | Credible | 1 | 9.75 % |
| | Not Credible | 0 | 8.44 % |
| | Not Judged | −1 or −2 | 81.81 % |

**Medical Intervention Efficacy Assessment** The assessors were asked to judge the medical intervention efficacy according to the content of the web document, and not what they believed is the correct information. For example, if a document mentions that a medical intervention helps, but the assessor knows that it does not — based on previous knowledge or acquired knowledge from assessing other documents — the document's medical intervention efficacy is still judged as Effective. Assessors were not required to have any medical knowledge to assess efficacy because their task does not include testing the correctness of the information presented. Treatment efficacy was only judged when the topical relevance of the document was Highly Relevant or Relevant.

*Efficacy* was labelled as one of the following:

- *Effective*: If the document states that the medical intervention is or can be an effective option to the health issue. If the document contains evidences for both the ineffective and effective directions, but it clearly supports the effective option over the ineffective one.
- *Inconclusive*: If the document contains evidence for both the ineffective and effective directions, but it does not clearly support one over the other. Or, it states that it is unknown whether or not the medical intervention helps. Or, it explicitly mentions the medical intervention, but does not provide any information on its efficacy, benefits, or disadvantages.
- *Ineffective*: If the document states that the medical intervention is ineffective or harmful. If the document contains evidence for both the ineffective and effective directions, but it clearly supports the ineffective option over the effective one;
- *No Information*: If the document does not state the health issue, but the assessor considered it relevant.

**Credibility** Understanding the purpose of a document should be the first step to judge credibility, thus the assessor's opinion of the purpose of the document and the credibility of information are fundamental in judging credibility. The idea of understanding the purpose of a website before judging its quality, determining the amount of Expertise, Authoritativeness, and Trustworthiness (E-A-T), is based on Google Search Quality Evaluator Guidelines [4].

[4]https://static.googleusercontent.com/media/guidelines.raterhub.com/en/searchqualityevaluatorguidelines.pdf Last visited: October 2019.

Credibility is only judged if the topical relevance of the document is Highly Relevant or Relevant. Furthermore, credibility should not be confused with topical relevance or efficacy and was judged independently of them, since a Not Credible document may be equally useful or helpful for a person making his/her decision.

*Credibility* was labelled as one of:

- *Credible*: Some criteria used to consider a document credible are: 1) if the document has a high level of E-A-T, 2) includes an author or a publishing institute expert in the field, 3) includes citations or references to credible sources such as universities, research/clinics, government websites, medical publications, and lab studies, 4) is hosted in a hospital/clinic or government website, or online newspaper with wide circulation, is well written, motivated and organized.
- *Not credible*: If the document is a mask for advertising or marketing purposes, is from a personal blog or a forum, or written by a non-expert person. If the document itself, or the hosting website, provides or claims that go against well-known medical consensus (e.g., smoking cigarettes does not cause cancer).

**Correctness Mapping** The correctness labels are computed based on a document's assessed efficacy by comparing the topic efficacy with the document efficacy. We consider a document *Correct* if the document label matches the document topic label. For example, consider Topic 8 in Figure 1, suggesting that melatonin can be used to reduce jet lag. Our truth is our interpretation of the systematic review done by the medical experts from Cochrane Review, which, in our opinion, concluded that melatonin is helpful. If a document for this topic was judged as Effective, i.e. the document claims that melatonin is helpful to reduce jet lag, then the document is considered correct, whereas a document judged as Ineffective or Inconclusive is considered not correct. Table 2 reports the mapping used to obtain the correctness labels on documents. Correctness labels for unjudged or non-relevant documents are not computed. Unjudged documents are assumed to be non-relevant, and non-relevant documents were not judged for efficacy or credibility.

Table 2: Mapping of the correctness values from the topic and document labels, 1 stands for correct, while 0 stands for incorrect.

| Topics \ Documents | Effective | Inconclusive | Ineffective | No Information |
|---|---|---|---|---|
| Helpful | 1 | 0 | 0 | 0 |
| Inconclusive | 0 | 1 | 0 | 0 |
| Not Helpful | 0 | 0 | 1 | 0 |

Documents that do not provide information on the efficacy of the medical intervention (i.e. judged as "No Information" by the assessor) are considered to be "Not Correct" since they do not provide any evidence to support or reject the claim included in the topic narrative. As shown in Table 1, the percentage of documents judged as No Information is particularly low, 0.38%, thus we do not expect them to affect the correctness labels significantly.

```
1 0 clueweb12-0000wb-03-01030 1 2 0
1 0 clueweb12-0000wb-47-24784 1 3 1
1 0 clueweb12-0000wb-54-11923 0 -1 -1
4 0 clueweb12-1902wb-14-21300 1 -2 0
```

(a) Original qrels

```
1 0 clueweb12-0000wb-03-01030 1 0 0
1 0 clueweb12-0000wb-47-24784 1 0 1
1 0 clueweb12-0000wb-54-11923 0 -1 -1
4 0 clueweb12-1902wb-14-21300 1 -2 0
```

(b) Correctness qrels

Figure 2: Sample of the original qrels as they were assessed by NIST (2a), and after the mapping to the correctness values (2b).

Finally, Figure 2 reports some sample documents from the qrels file before and after the mapping to the correctness labels. The new qrels are formatted as the original qrels from NIST, except for the efficacy value — second to last column — which is replaced by the correctness value. For example, document `clueweb12-0000wb-03-01030` was assessed as Inconclusive, but topic 1 was assessed as Helpful,

therefore the document is Not Correct. Similarly, document `clueweb12-0000wb-47-24784` was assessed as effective for topic 1, thus it is considered Not Correct. The documents `clueweb12-0000wb-54-11923` and `clueweb12-1902wb-14-21300` were not judged for efficacy, therefore they result as not judged also for correctness.

## 2.3 Evaluation Measures

We evaluated Task 1 with two different approaches: firstly as a standard ad-hoc retrieval task, i.e. we considered just topical relevance, and secondly we considered all the aspects: relevance, correctness, and credibility. The purpose of using the two approaches to evaluation was to investigate the extent to which merely evaluating retrieval quality by relevance can fail to measure other important aspects of retrieval quality.

### 2.3.1 Ad-hoc Retrieval Evaluation

For the standard ad-hoc retrieval evaluation, we used Average Precision (AP) [5] and Normalized Discounted Cumulated Gain (nDCG) [9] with a cut-off at 10. We used `trec_eval`[5] to compute AP and nDCG.

### 2.3.2 Multi-aspect Evaluation

To evaluate the runs with respect to the three aspects, we selected two measures from Lioma et al. [13], namely Convex Aggregating Measure (CAM) and Normalized Local Rank Error (NLRE). Note that the efficacy labels for documents and topics were not used for the evaluation, but we considered correctness instead. Both CAM and NLRE are originally defined for two aspects, topical relevance and credibility, therefore we extended the definition of those measures to deal with correctness as well. Note that, unjudged documents are considered to be Non-relevant, Not Correct, and Not Credible by both these measures.

**Convex Aggregating Measure (CAM)** Let $r$ be a ranked list of documents with multi-aspect labels, *Convex Aggregating Measure (CAM)* is defined as the convex sum of the $\mathcal{M}$ scores computed with respect to each aspect individually:

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

where $\mathcal{M}_{rel}$, $\mathcal{M}_{cor}$, and $\mathcal{M}_{cre}$ denote respectively any valid relevance, correctness, and credibility evaluation measure, and $\lambda_{rel} + \lambda_{corr} + \lambda_{cre} = 1$ are non negative parameters controlling the impact of the individual relevance, correctness and credibility measures in the overall computation.

We instantiated CAM with $\lambda_{\#} = 1/3$, thus assigning the same weight to each aspect. As evaluation measure we used nDCG for each individual aspect, i.e. $\mathcal{M}_{rel}$ is the standard nDCG computed with respect to relevance, $\mathcal{M}_{cor}$ is nDCG computed with respect to the correctness labels, while $\mathcal{M}_{cre}$ is nDCG computed with respect to the credibility labels. We did not use F-1 or G-measure for credibility, as proposed by Lioma et al [13], since we wanted to account for the rank position of credible and not credible documents as well.

**Normalized Local Rank Error (NLRE)** *Normalized Local Rank Error (NLRE)* [13] accounts for the error computed with respect to three additional ideal re-rankings independent of each other: one by relevance only, one by correctness only, and one by credibility only. Given an input ranked list $r$, Normalized Local Rank Error (NLRE) takes adjacent pairs of documents in $r$ and checks for errors in $r$ compared to the three ideal re-rankings. Let $d_i$ be the document at rank position $i$ in the ranked list $r$, then let $r_{rel}[d_i]$ be the rank position of $d_i$ in the ideal ranked list computed with respect to relevance only, similarly $r_{cor}[d_i]$ is the rank position of $d_i$ in the ideal ranked list by correctness and $r_{cre}[d_i]$ is the rank position of $d_i$ in the ideal ranked list by credibility.

[5] https://trec.nist.gov/trec_eval/

The relevance, correctness and credibility errors for $d_i$, namely $\epsilon_{rel}[d_i]$, $\epsilon_{cor}[d_i]$, and $\epsilon_{cre}[d_i]$ are defined as follows:

$$\begin{aligned}\epsilon_{rel}[d_i] &= \max\{0, r_{rel}[d_i] - r_{rel}[d_{i+1}]\} \\ \epsilon_{cor}[d_i] &= \max\{0, r_{cor}[d_i] - r_{cor}[d_{i+1}]\} \\ \epsilon_{cre}[d_i] &= \max\{0, r_{cre}[d_i] - r_{cre}[d_{i+1}]\}\end{aligned} \tag{2}$$

For example, given two documents $d_i$ and $d_{i+1}$ in $r$, the relevance error is greater than zero if the document $d_i$ is ranked after $d_{i+1}$ in the ideal re-ranked list computed by relevance only.

Let $n$ be the total number of documents in the ranked list. We define the Local Rank Error (LRE) evaluation measure as LRE $= 0$ if $n = 1$ and otherwise:

$$\text{LRE} = \sum_{i=1}^{n-1} \frac{1}{\log_2(1+i)} \Big( (\mu + \epsilon_{rel}[d_i])(\nu + \epsilon_{cor}[d_i])(\xi + \epsilon_{cre}[d_i]) - \mu\nu\xi \Big) \tag{3}$$

where $\epsilon_{\#}[d_i]$ are the errors as defined in Equation 2, and $\mu$, $\nu$, $\xi$ are non negative real numbers controlling how much each aspect should be penalised. As for CAM, we assigned the same weight to each aspect, thus setting $\mu = \nu = \xi = 1/3$.

Finally, because Equation 3 is large for bad rankings and small for good rankings, we invert and normalize it as follows:

$$\text{NLRE} = 1 - \frac{\text{NLRE}}{C_{\text{LRE}}} \tag{4}$$

where $C_{\text{LRE}}$ is the normalization constant, defined as:

$$C_{\text{LRE}} = \sum_{j=0}^{\lfloor \frac{n}{2} - 1 \rfloor} \frac{(n-2j-1)^3 + (\mu+\nu+\xi)(n-2j-1)}{1 + \log_2(1+j)} \tag{5}$$

Equation 5 ensures that NLRE$/C_{\text{LRE}} \leq 1$. $C_{\text{LRE}}$ computes the maximum possible error attainable, i.e. rankings that produce the largest possible relevance, correctness and credibility errors. NLRE is 1 if no errors of any kind occur, since in this case LRE is 0.

**MM framework** Along with CAM and NLRE, we also computed nDCG@10 and AP using the MM framework for multidimensional relevance evaluation [15], as an alternative approach to incorporate correctness and credibility in the evaluation alongside topical relevance. In the MM framework, the evaluation scores for each aspect of relevance (topical relevance, correctness, credibility) is calculated separately according to a selected evaluation measure (e.g., nDCG@10, AP), and then these are combined into a unique effectiveness measure using the weighted harmonic mean. The weighted harmonic mean is particularly sensitive to a single lower-than-average value, thus rewarding systems that are consistently more effective across all aspects of relevance. (The same intuition is used to combine recall and precision in the widely used $F$-measure.) Given an evaluation measure $\mathcal{M}$, we apply the measure to evaluate a ranked list of documents $r_\delta$ which have been labeled with respect to aspect $\delta$ (i.e., we compute $\mathcal{M}(r_\delta)$). Then, to compute $MM_{\mathcal{M}}$, all $\mathcal{M}(r_\delta)$ are combined for each aspect using the harmonic mean, where each aspect is weighted according to a preferential weight $w_\delta$ assigned to each aspect, as it was for $\lambda_\#$ for CAM. Formally:

$$MM_{\mathcal{M}} = \left( \frac{\sum_{\delta=1}^{n} w_\delta \cdot \mathcal{M}(r_\delta)^{-1}}{\sum_{\delta=1}^{n} w_\delta} \right)^{-1} = \frac{\sum_{\delta=1}^{n} w_\delta}{\sum_{\delta=1}^{n} \frac{w_\delta}{\mathcal{M}(r_\delta)}} \tag{6}$$

We set $w_\delta$ to $1/3$ across all aspects, and compute the MM variants for AP (MM(AP)) and nDCG@10 (MM(nDCG@10)). Note that MM-based evaluation results were not distributed to participants at the time of writing their notebook papers.

# 3 Participation and Experimental Evaluation

We received 32 runs from 4 groups: University of Waterloo (UWaterlooMDS), University of Queensland (UE IELab), Chinese Academy of Sciences (ICTNET), and Bauhaus-Universität Weimar (Webis). Table 3 reports the runs submitted by each group. UWaterlooMDS was the only group that generated both automatic and manual runs – all other groups generated automatic runs. IElab was the only group that considered both information in the query field and in the other portion of the topic to generate runs – all other groups only used the query field. A brief summary of each group's submissions:

**Webis** The Webis group [4] manipulated Elasticsearch's BM25F initial ranking based on the credibility of documents' web hostnames' domains. The result is then re-ranked using an axiomatic approach that captures argumentativeness and information credibility.

**ICTNET** The ICTNET group [6] used Terrier's BM25 as their method of retrieval. The group also considered other retrieval methods.

**UQ IElab** The UQ IElab group [11] employed query expansion methods using knowledge-bases (e.g. Wikipedia) to capture medical vocabulary from the topic fields. The underlying retrieval method used is Elasticsearch's BM25F.

**UWaterlooMDS** The UWaterlooMDS group [2] submitted manual and automatic runs. For manual runs, the group used HiCAL [1], an open-source high-recall retrieval system, to retrieve and manually judge documents. The manually judged documents are then used to re-rank documents from a baseline BM25 ranking obtained using Ansereni[6]. For automatic runs, the team built a credibility classifier trained on an annotated corpus prepared using HiCAL for finding non-credible documents. The automatic runs combined spam and credibility classifier scores to modify a BM25 baseline run.

## 3.1 Results

**Ad-hoc Retrieval Evaluation** Table 4 reports the AP and nDCG scores of the submitted runs when only topical relevance is considered. Measures are averaged across topics. Tables 6 and 7 report the statistical analysis of differences in effectiveness scores among the top runs from each group. Furthermore, Figure 3 and Figure 4 provide a per-topic analysis for AP and nDCG@10.

When AP is considered, runs from UWaterlooMDS (University of Waterloo) are consistently better than those from other groups, as shown in Figure 3. Indeed, the first, second and third best runs are respectively `UWatMDSBM25_HC3`, `UWatMDSBM25_HC1`, `UWatMDSBM25_HC2`. In addition, the best run from UWaterlooMDS with respect to AP is statistically significantly different from the best run of the runner up team, IElab (paired two tails t-test with Bonferroni correction, see Table 6). Furthermore, the best run from UWaterlooMDS has a relative improvement of 0.2322 over the best run from IElab (UQ), 1.7314 over the best run from Webis (Bauhaus-Universität Weimar), and 110.0270 over the best run from ICTNET (Chinese Academy of Sciences).

Similar findings are obtained when nDCG@10 is considered, with UWaterlooMDS providing the best two runs for this measure, although the actual runs that achieve the highest nDCG@10 values (`UWatMDS_BMF_S30` and `UWaterMDS_BM25`) are different from those that obtain the highest AP values. Unlike AP, the best run from UWaterlooMDS with respect to nDCG@10 is not statistically significantly different from the best run of the runner up team, IElab (see Table 7). The relative improvement of the best run from UWaterlooMDS over the best run from IElab is 0.0356, while it is 0.9516 for Webis, and 13.7493 for ICTNET.

Note that the improvement of UWaterlooMDS over the other teams is less pronounced when considering nDCG@10 compared to AP. This suggests that UWaterlooMDS best runs are better than the others at ranking relevant documents earlier across the *entirety* of the ranking; while when just the first 10 rank positions are considered, runs from UWaterlooMDS and IElab have comparable effectiveness. It is interesting to note that the UWaterlooMDS runs that are best for AP are not so when considering the number of relevant

[6] https://github.com/castorini/anserini

Table 3: Groups participating in TREC 2019 Decision Track and submitted runs. Runs are classified as either automatic (auto) or manual (man), and ad either based on the query field only (query), or also on other fields in the topic (other).

| Group | Affiliation | # Submissions | Runs | Type | Field Used |
|---|---|---|---|---|---|
| ICTNET | Chinese Academy of Sciences | 2 | ICTNETv1BM25 | auto | query |
| | | | ICTNETv2BM25 | auto | query |
| IElab | University of Queensland | 10 | IELAB01_ori_q | auto | query |
| | | | IELAB02_ori_d | auto | other |
| | | | IELAB03_umls_d | auto | other |
| | | | IELAB04_umls_n | auto | other |
| | | | IELAB05_xChv_q | auto | query |
| | | | IELAB06_xChv_d | auto | other |
| | | | IELAB07_xWiki_q | auto | query |
| | | | IELAB08_xWiki_d | auto | other |
| | | | IELAB09_xCW_q | auto | query |
| | | | IELAB10_xCW_d | auto | other |
| UWaterlooMDS | University of Waterloo | 10 | UWatMDSBM25_HC1 | man | query |
| | | | UWatMDSBM25_HC2 | man | query |
| | | | UWatMDSBM25_HC3 | man | query |
| | | | UWatMDS_BM25_Z | auto | query |
| | | | UWatMDS_BM25_ZS | auto | query |
| | | | UWatMDS_BMF_C90 | auto | query |
| | | | UWatMDS_BMF_C95 | auto | query |
| | | | UWatMDS_BMF_S30 | auto | query |
| | | | UWatMDS_BMZBS10 | auto | query |
| | | | UWaterMDS_BM25 | auto | query |
| Webis | Bauhaus-Universität Weimar | 10 | webisMAll1 | auto | query |
| | | | webisMMajority1 | auto | query |
| | | | webisMSame1 | auto | query |
| | | | webisMSame2 | auto | query |
| | | | webisMSame3 | auto | query |
| | | | webisMSame4 | auto | query |
| | | | webisMSame5 | auto | query |
| | | | webisTAll1 | auto | query |
| | | | webisTMajority1 | auto | query |
| | | | webisTSame1 | auto | query |

documents retrieved – in fact IELAB01_ori_q has a better recall, retrieving 70.66 relevant documents per queries on average, compared to the 66.76 of UWaterMDSBM25_HC3. Nevertheless, the best two runs in terms of number of relevant documents retrieved are from UWaterlooMDS (UWaterMDS_BM25 and UWaterMDS_BM25_Z retrieve 72.86 and 71.92 relevant documents per query, on average).

**Multi-aspect Evaluation** Table 5 reports the evaluation results when all aspects are considered (relevance, correctness, credibility) and CAM and NLRE are used. The measures are averaged across topics. Tables 8 and 9 report the statistical analysis of differences in effectiveness scores among the top runs from each group. Furthermore, Figure 5 provides a per-topic analysis for CAM.

The findings for the multi-aspect evaluation are similar to those when only topical relevance is considered (ad-hoc evaluation). When CAM is considered, in fact, the most effective runs remain those from UWaterlooMDS (University of Waterloo): the first, second and third best runs are respectively UWaterMDS_BM25, UWatMDS_BM25_Z, UWatMDSBM25_HC3. The best run from UWaterlooMDS has a relative improvement of 0.0517

Table 4: Evaluation results where only relevance is considered. The average score over the topics is reported. The best scores for each team are in bold. The overall first, second and third runs are denoted by $\star\star\star$, $\star\star$, and $\star$ respectively.

| Run Name | AP | nDCG@10 |
|---|---|---|
| ICTNETv1BM25 | 0.0000 | 0.0007 |
| ICTNETv2BM25 | **0.0037** | **0.0339** |
| IELAB01_ori_q | **0.3334** | **0.4828**$^{\star}$ |
| IELAB02_ori_d | 0.2082 | 0.3525 |
| IELAB03_umls_d | 0.2723 | 0.3948 |
| IELAB04_umls_n | 0.2387 | 0.3565 |
| IELAB05_xChv_q | 0.2642 | 0.4125 |
| IELAB06_xChv_d | 0.2613 | 0.3906 |
| IELAB07_xWiki_q | 0.3129 | 0.4651 |
| IELAB08_xWiki_d | 0.2718 | 0.3936 |
| IELAB09_xCW_q | 0.2568 | 0.4065 |
| IELAB10_xCW_d | 0.2611 | 0.3898 |
| UWatMDSBM25_HC1 | 0.4027$^{\star\star}$ | 0.4504 |
| UWatMDSBM25_HC2 | 0.3911$^{\star}$ | 0.4504 |
| UWatMDSBM25_HC3 | **0.4108**$^{\star\star\star}$ | 0.4504 |
| UWatMDS_BM25_Z | 0.3448 | 0.4430 |
| UWatMDS_BM25_ZS | 0.3105 | 0.4302 |
| UWatMDS_BMF_C90 | 0.1562 | 0.4249 |
| UWatMDS_BMF_C95 | 0.1699 | 0.4450 |
| UWatMDS_BMF_S30 | 0.2855 | **0.5000**$^{\star\star\star}$ |
| UWatMDS_BMZBS10 | 0.2827 | 0.3921 |
| UWaterMDS_BM25 | 0.3764 | 0.4986$^{\star\star}$ |
| webisMAll1 | **0.1504** | **0.2562** |
| webisMMajority1 | 0.1475 | 0.2395 |
| webisMSame1 | 0.1402 | 0.1641 |
| webisMSame2 | 0.1382 | 0.1510 |
| webisMSame3 | 0.1417 | 0.1481 |
| webisMSame4 | 0.1354 | 0.1149 |
| webisMSame5 | 0.1349 | 0.1162 |
| webisTAll1 | 0.1495 | 0.2512 |
| webisTMajority1 | 0.1491 | 0.2498 |
| webisTSame1 | 0.1373 | 0.1447 |

Table 5: Evaluation results where all the aspects are considered. The average score over the topics is reported. The best scores for each team are in bold. The overall first, second and third runs are denoted by $\star\star\star$, $\star\star$, and $\star$ respectively.

| Run Name | CAM | NLRE |
|---|---|---|
| ICTNETv1BM25 | 0.0001 | **1.0000**$^{\star\star\star}$ |
| ICTNETv2BM25 | **0.0085** | **1.0000**$^{\star\star\star}$ |
| IELAB01_ori_q | **0.5208** | 0.9960 |
| IELAB02_ori_d | 0.3872 | 0.9959 |
| IELAB03_umls_d | 0.4547 | **0.9966** |
| IELAB04_umls_n | 0.4234 | 0.9957 |
| IELAB05_xChv_q | 0.4665 | 0.9963 |
| IELAB06_xChv_d | 0.4493 | 0.9959 |
| IELAB07_xWiki_q | 0.5084 | 0.9962 |
| IELAB08_xWiki_d | 0.4547 | 0.9963 |
| IELAB09_xCW_q | 0.4621 | 0.9963 |
| IELAB10_xCW_d | 0.4491 | 0.9961 |
| UWatMDSBM25_HC1 | 0.5360 | 0.9972 |
| UWatMDSBM25_HC2 | 0.5336 | 0.9976 |
| UWatMDSBM25_HC3 | 0.5386$^{\star}$ | 0.9983$^{\star}$ |
| UWatMDS_BM25_Z | 0.5467$^{\star\star}$ | 0.9968 |
| UWatMDS_BM25_ZS | 0.5096 | 0.9969 |
| UWatMDS_BMF_C90 | 0.3089 | **0.9991**$^{\star\star}$ |
| UWatMDS_BMF_C95 | 0.3339 | **0.9991**$^{\star\star}$ |
| UWatMDS_BMF_S30 | 0.4560 | 0.9978 |
| UWatMDS_BMZBS10 | 0.4918 | 0.9971 |
| UWaterMDS_BM25 | **0.5477**$^{\star\star\star}$ | 0.9958 |
| webisMAll1 | **0.3773** | **0.9940** |
| webisMMajority1 | 0.3711 | 0.9940 |
| webisMSame1 | 0.3479 | 0.9940 |
| webisMSame2 | 0.3468 | 0.9940 |
| webisMSame3 | 0.3518 | 0.9940 |
| webisMSame4 | 0.3419 | 0.9940 |
| webisMSame5 | 0.3436 | 0.9940 |
| webisTAll1 | 0.3758 | 0.9940 |
| webisTMajority1 | 0.3727 | 0.9940 |
| webisTSame1 | 0.3462 | 0.9940 |

Table 6: Statistical significance analysis (p-values obtained with paired two tails t-test, with Bonferroni correction) on AP scores for the best runs submitted by each team.

| | ICTNETv2BM25 | IELAB01_ori_q | UWatMDSBM25_HC3 |
|---|---|---|---|
| IELAB01_ori_q | <2e-16 | - | - |
| UWatMDSBM25_HC3 | <2e-16 | 0.0012 | - |
| webisMAll1 | 5.6e-11 | 3.2e-15 | <2e-16 |

Table 7: Statistical significance analysis (p-values obtained with paired two tailed t-test, with Bonferroni correction) on nDCG@10 scores for the best runs submitted by each team.

| | ICTNETv2BM25 | IELAB01_ori_q | UWatMDS_BMF_S30 |
|---|---|---|---|
| IELAB01_ori_q | 3.0e-15 | - | - |
| UWatMDS_BMF_S30 | < 2e-16 | 1 | - |
| webisMAll1 | 1.0e-07 | 1.9e-07 | 5.4e-09 |

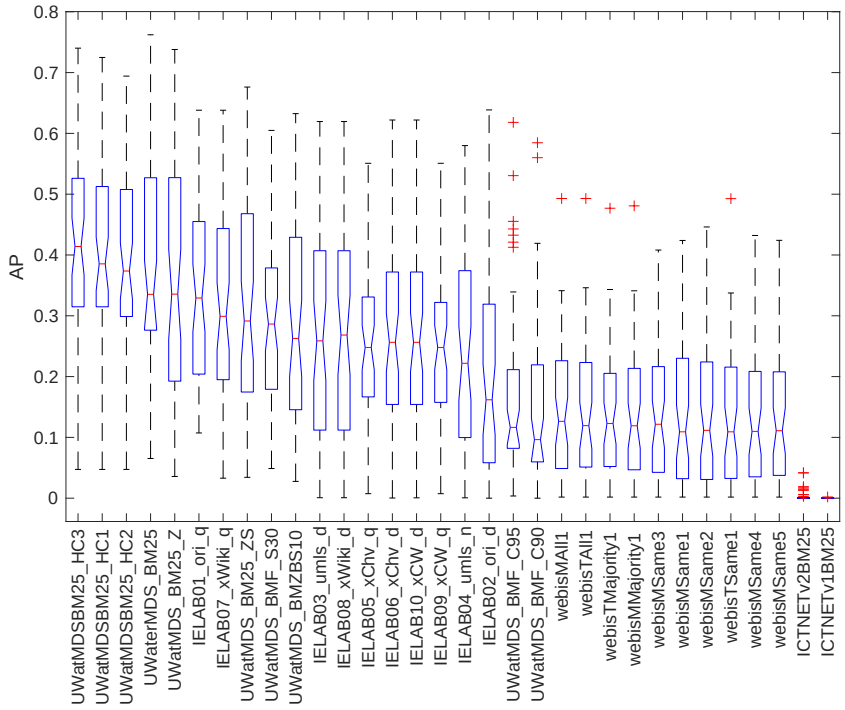


Figure 3: Relevance based evaluation: box-plot of AP scores over the 50 topics. The runs are sorted by descending average score.

over the best run from IElab, and 0.4516 and 63.4353 over the best run from Webis and ICTNET, respectively.

Note that CAM is defined as the average of nDCG, computed separately with respect to each different aspect. If nDCG was to be computed on the whole ranking with respect to topical relevance only, the results obtained using CAM will be somehow aligned with those obtained using nDCG. Indeed the first and second best runs with respect to nDCG@1000 coincide with the first and second run with respect to CAM.

Furthermore, the scores computed with CAM are generally lower in terms of absolute values than the scores computed with nDCG@1000. This shows that considering both correctness and credibility affects

Table 8: Statistical significance analysis (p-values obtained with paired two tailed t-test, with Bonferroni correction) on CAM scores for the best runs submitted by each team.

| | ICTNETv2BM25 | IELAB01_ori_q | UWaterMDS_BM25 |
|---|---|---|---|
| IELAB01_ori_q | < 2e-16 | - | - |
| UWaterMDS_BM25 | < 2e-16 | 0.082 | - |
| webisMAll1 | < 2e-16 | 3.9e-10 | 5.1e-11 |

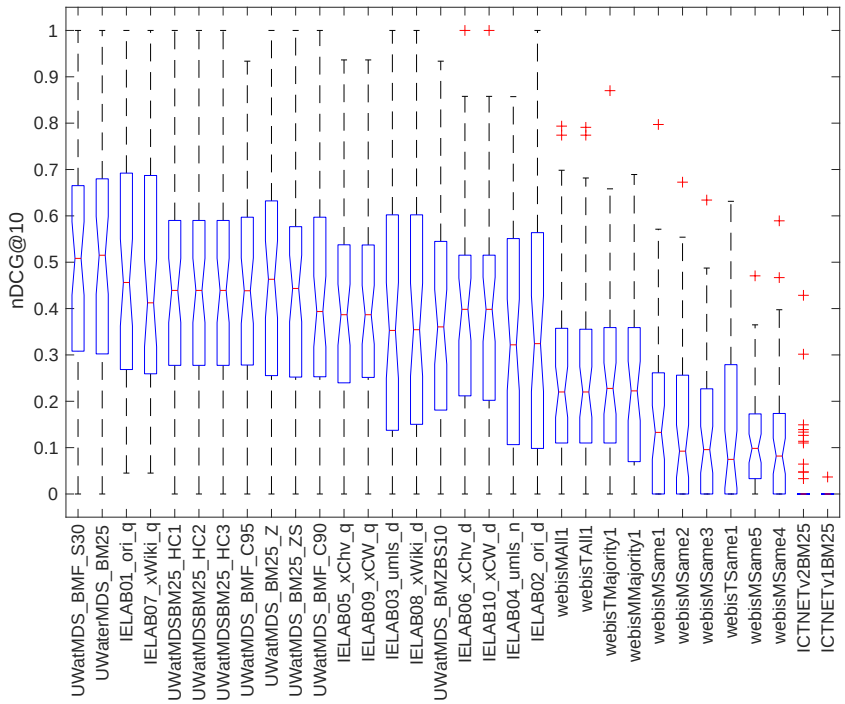


Figure 4: Relevance based evaluation: box-plot of nDCG@10 scores over the 50 topics. The runs are sorted by descending average score.

the evaluation, sometimes also by changing the global ranking of systems, as for example `UWatMDSBM25_HC3` which is the third best run for CAM, while it is ranked after `IELAB01_ori_q` for nDCG@1000.

Table 5 reports NLRE average scores and Figure 6 shows the box-plot with the per-topic analysis for NLRE. As shown in Figure 6, NLRE scores are surprisingly close to 1, the reason is detailed in Section 3.2.

With respect to NLRE, the best group is ICTNET, followed by UWaterlooMDS. Note that for NLRE the ranking of groups is extremely different from the ranking obtained with all the other measures. This is due to the definition of NLRE, which exploits the idea of computing the error between the ranking and the ideal re-ranking, which is different from both AP and nDCG.

Finally, the evaluation results obtained using the MM framework is reported in Figures 7 and 8 for AP and nDCG@10, respectively. The trends observed for MM(nDCG@10) are similar to those obtained with CAM: this is not surprising, as both measures are based on the interpolation of nDCG@10 computed separately for each aspect. However, we observe that the use of the harmonic mean in MM, rather than the arithmetic mean as in CAM does provide some key differences. For example, run `IELAB01_ori_q` (the best from the IElab group) ranks 6th according to CAM, but it only ranks 11th according to MM(nDCG@10) (although

Table 9: Statistical significance analysis (p-values obtained with paired two tailed t-test, with Bonferroni correction) on NLRE scores for the best runs submitted by each team. Note, we ignored run ICTNETv2BM25 as it was equivalent to run ICTNETv1BM25, reported here.

| | ICTNETv1BM25 | IELAB03_umls_d | UWatMDS_BMF_C90 | UWatMDS_BMF_C95 |
|---|---|---|---|---|
| IELAB03_umls_d | 0.0076 | - | - | - |
| UWatMDS_BMF_C90 | 0.1589 | 0.0110 | - | - |
| UWatMDS_BMF_C95 | 0.0846 | 0.0182 | 1.0000 | - |
| webisMAll1 | 0.0055 | 0.2437 | 0.0108 | 0.0101 |

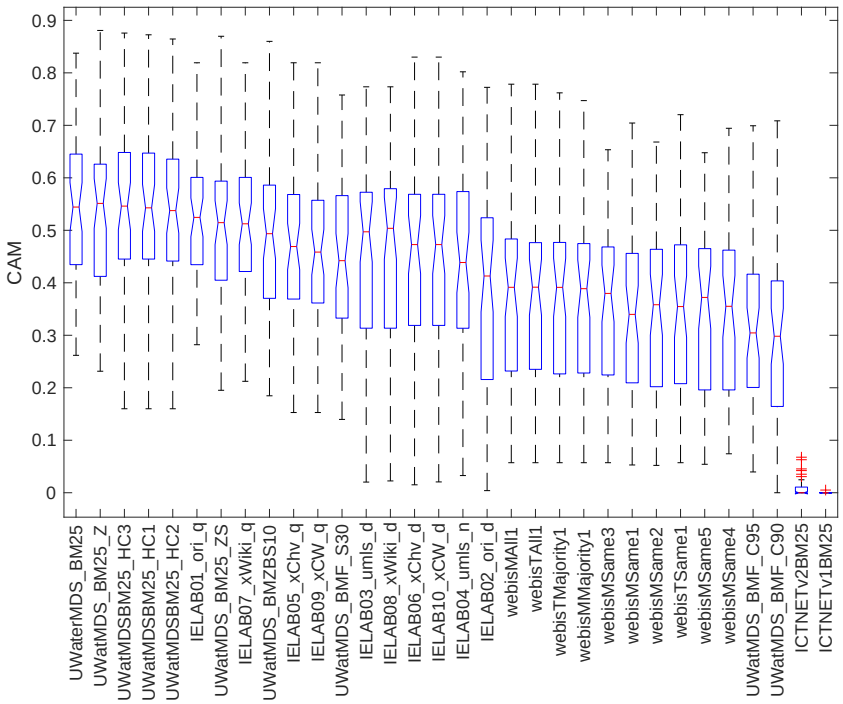


Figure 5: Multi-aspect evaluation: box-plot of CAM scores over the 50 topics. The runs are sorted by descending average score.

it still is the best run for that group). This difference is due to the harmonic mean punishing the fact that the run perform particularly worse than the other tuns from UWaterlooMDS in a specific relevance aspect, while it does perform better in the other two aspects – while the runs from UWaterlooMDS perform more consistently across all aspects. We also observe that when using MM(nDCG@10) for evaluation, systems are mostly indistinguishable (note the whiskers in Figure 8), except for the runs from ICTNET and two of the UWaterlooMDS runs.

## 3.2 Discussion on Multi-aspect Evaluation Measures

As mentioned in Section 1, one of the goal of the Decision Track is to investigate possible strategies to develop new offline evaluation methods able to account for multiple aspects simultaneously. To this end, we can draw two main conclusions from Table 5: (1) CAM is more reliable than NLRE to some extent, but it is closely bound to nDCG due to its definition; (2) NLRE has some limitations that prevent a proper understanding of the effectiveness performance.

Table 5 and Figure 6 clearly show two main limitations of NLRE: all the scores are close or equal to the perfect score, suggesting that the runs perform well even if the other measures show that this is not the case. That is, the measure cannot discriminate among runs: for example, all the runs from the Webis group have the same NLRE score. We identified the reasons behind these limitations in two main issues: NLRE considers the ideal ranking as a re-ranking of the input ranking; the normalization constant assumes very large values for three aspects and when the whole ranking is considered.

First, when NLRE computes the error with respect to each aspect, it does not consider an ideal ranking on the whole collection, but simply a re-ranking of the input ranking. Theoretically, the measure would work if every run sorts the whole collection or if every run retrieves all the relevant, correct, and credible documents. However, both the assumptions are often not satisfied in a real world scenario, as it is the case with this TREC task where just the first 1000 documents are considered. In other words, this means that

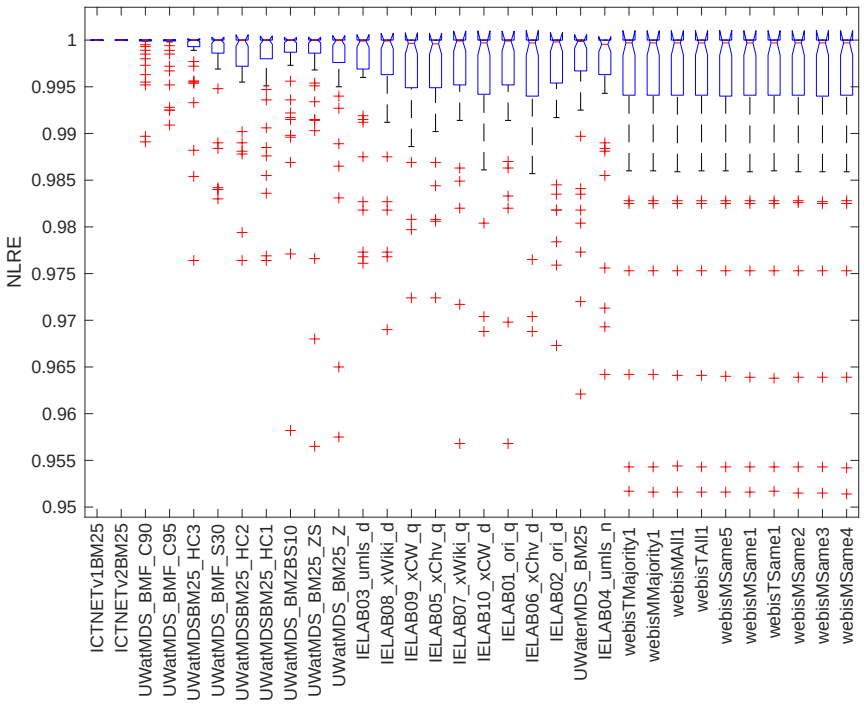


Figure 6: Multi-aspect evaluation: box-plot of NLRE scores over the 50 topics. The runs are sorted by descending average score.

NLRE does not make use of the recall base.

The practical consequence of the definition of NLRE upon the ideal re-rankings, is that if a ranked list does not retrieve any relevant, correct or credible document, then LRE is equal to 0 because there are no errors in the input ranking, being all the documents correctly sorted. This effect is particularly evident from the first two lines of Table 5 representing the first two runs in Figure 6, where both `ICTNETv1BM25` and `ICTNETv2BM25` have NLRE average score equal to 1, even if CAM score is close to 0, meaning that just few relevant documents are retrieved.

Second, NLRE was originally tested on a sample of runs with two aspects, where just the first 5 documents where considered, i.e. only NLRE@5 was computed [13]. In that case the value of the normalization constant is $C_{\text{LRE}} = 20$. The results reported in Table 5 consider the whole ranked list instead. In this case, the normalization constant, which computes the maximum possible error, is $C_{LRE} \sim 20 * 10^9$. Therefore, LRE scores are divided by a large constant, which means that $\text{LRE}/C_{\text{LRE}} \to 0$ and consequently NLRE $\to 1$. This explains why all the scores in Table 5 and in Figure 6 are particularly close to 1 and also why the measure does not effectively discriminate among different runs, as is the case for example with the runs by Webis.

While CAM is more reliable than NLRE in evaluating the true performance of runs, it has some potential issues. Since one of the track's goals is to devise search technologies that promote correct information over incorrect information, it is reasonable to claim that a non-relevant document is preferable to a relevant document with incorrect and potentially harmful information. As previous work has shown, such documents can influence users to make incorrect or harmful decisions, and search engines should avoid presenting incorrect documents to users [16]. Unfortunately, the definition of CAM does not take such cases into consideration.

The aggregation of the same measure using different aspects (e.g. relevance, correctness, and credibility) in CAM also means that returning a non-relevant document would penalize the performance of a run in terms of both correctness and credibility, when clearly it should penalize for relevance only.

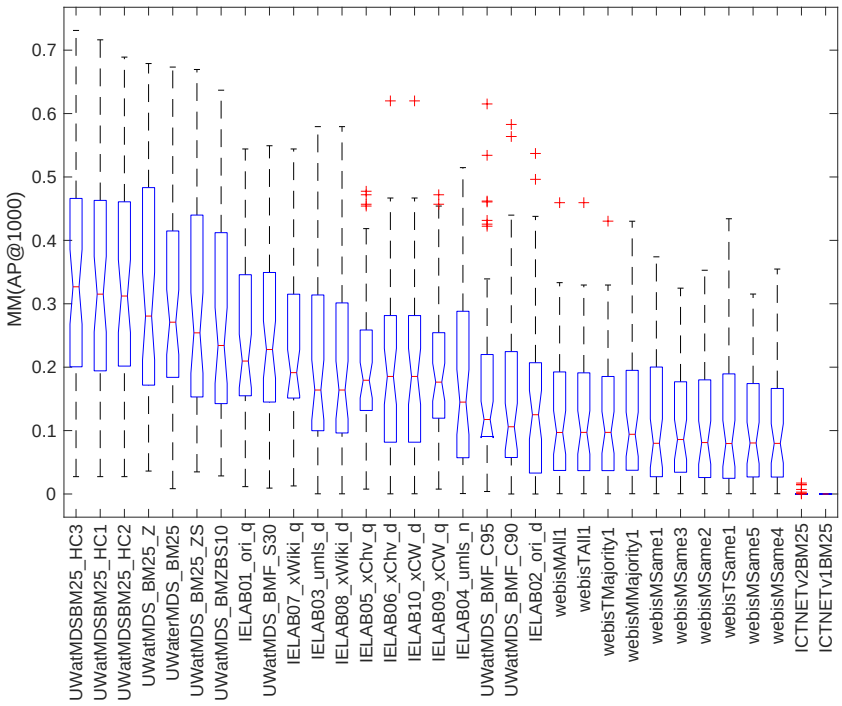


Figure 7: Multi-aspect evaluation: box-plot of MM(AP@1000) scores over the 50 topics. The runs are sorted by descending average score.

Another observation of the pitfalls of CAM can be inferred from Figure 5. CAM scores `UWaterlooMDS_BM25` run as the best performing run, while the analysis from the UWaterlooMDS team in Abualsaud et al. [2] shows that some of their retrieval methods were able to push credible or correct documents to the top of the ranking relative to their original position in their baseline run `UWaterlooMDS_BM25`, which may indicate that CAM may not capture performance as intended.

### 3.3 Incomplete assessments: pool coverage

As it is common in information retrieval evaluation, not all documents retrieved by the submitted systems could be assessed by the NIST judges, due to budget constraints. The depth@k pooling method was used to select documents from the submitted retrieval runs to assess. All the submitted systems contributed to the pool. The depth $k$ was initially set at 75, but due to time limitations this was reduced to 60 for some topics. Table 10 reports the average coverage of the relevance assessments for each of the submitted runs, i.e. the percentage of documents in the runs for which a relevance assessment has been recorded.

We furthermore study whether relevance assessment coverage and ad-hoc evaluation performance are correlated: for example, one may hypothesise that the top performing runs do so because most of their retrieved documents have been assessed, while the lower performing runs present more missing judgements. We found this not to be the case; specifically, both AP and nDCG@10 present no correlation with the coverage of the assessments (Pearson's correlation 0.1748 and 0.0283, respectively, and $\tau_{AP}$ correlation [25] 0.0225 and -0.0547, respectively).

NIST ran out of assessing budget before Topic 14, and thus Topic 14 is excluded.

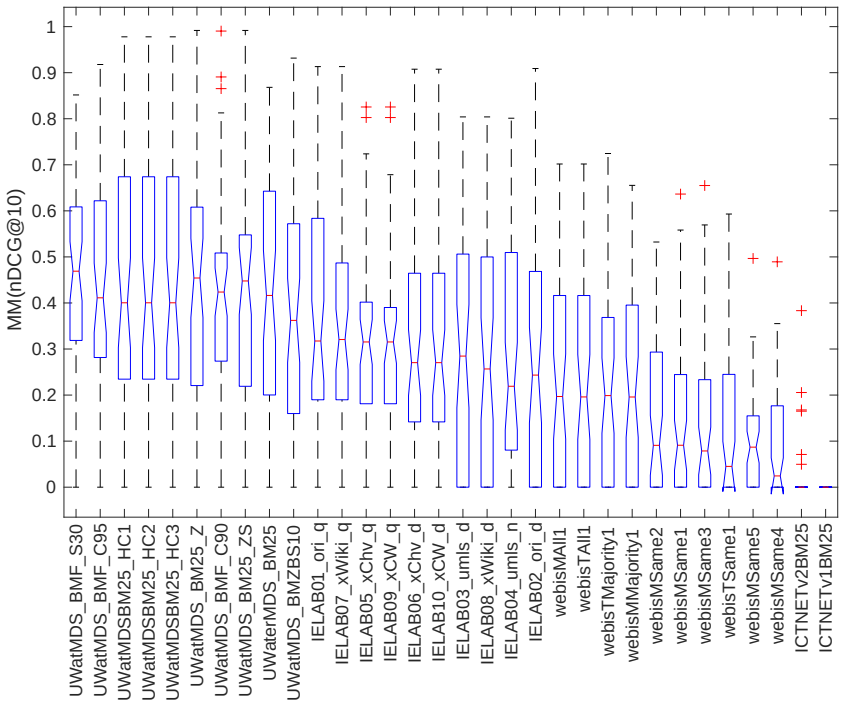


Figure 8: Multi-aspect evaluation: box-plot of MM(nDCG@10) scores over the 50 topics. The runs are sorted by descending average score.

Table 10: For each submitted run, we report the portion of the result list that has a relevance assessment (percentage), averaged across topics.

| Run Name | Avg. Coverage | Run Name | Avg. Coverage |
|---|---|---|---|
| webisMAll1 | 27.20% | UWatMDS_BM25_ZS | 21.99% |
| webisMMajority1 | 27.20% | IELAB06_xChv_d | 21.86% |
| webisMSame1 | 27.20% | IELAB10_xCW_d | 21.86% |
| webisMSame2 | 27.20% | IELAB05_xChv_q | 21.77% |
| webisMSame3 | 27.20% | IELAB09_xCW_q | 21.62% |
| webisMSame4 | 27.20% | UWatMDS_BMZBS10 | 21.53% |
| webisMSame5 | 27.20% | UWatMDSBM25_HC1 | 21.36% |
| webisTAll1 | 27.20% | UWatMDSBM25_HC2 | 21.36% |
| webisTMajority1 | 27.20% | UWatMDSBM25_HC3 | 21.36% |
| webisTSame1 | 27.20% | IELAB04_umls_n | 19.79% |
| IELAB01_ori_q | 24.29% | UWatMDS_BMF_S30 | 18.50% |
| UWatMDS_BM25_Z | 24.02% | IELAB02_ori_d | 18.09% |
| UWaterMDS_BM25 | 23.91% | UWatMDS_BMF_C95 | 13.30% |
| IELAB07_xWiki_q | 23.88% | UWatMDS_BMF_C90 | 12.02% |
| IELAB08_xWiki_d | 22.84% | ICTNETv2BM25 | 10.75% |
| IELAB03_umls_d | 22.83% | ICTNETv1BM25 | 6.15% |

# 4 Conclusions and Future Directions

The TREC 2019 Decision Track received 32 submissions from 4 groups: ICTNET, IELab, UWaterlooMDS and Webis. The empirical evaluation of the submitted runs that relies on relevance based measures, i.e. AP and nDCG@10, shows that the best performing group is UWaterlooMDS, followed by IElab and Webis. The evaluation considering all the aspects — (topical) relevance, correctness and credibility — is performed with respect to CAM and NLRE [13]. CAM is consistent with the results reported by the relevance based evaluation, with UWaterlooMDS being the leading group followed by IElab and Webis. NLRE results show a different evaluation perspective; however these results are affected by some biases due to the definition of NLRE. This further shows the necessity of well defined measures, able to account for multiple aspects simultaneously and without bias. Therefore, one of the goals of the track for next year will be to design new measures able to overcome the pitfalls shown by NLRE.

In 2020, the track will change its name to the TREC Health Misinformation Track. The organizers decided to change the name of the track to reflect its focus on health search and the effect of incorrect information on searcher decisions. The goals of the track remain the same, but we hope the new name will better communicate the track's tasks. In addition to changing the name of the track, the track organizers will change with Christina Lioma stepping down and Charles Clarke (University of Waterloo) joining.

The TREC Health Misinformation Track will have three tasks in 2020. The first task will be a repeat of the 2019 retrieval task but with a new set of search topics. The second task will be new and will be a recall task to find all health misinformation for each of the search topics. The third task, also new, will be to design a new offline measure to predict the decision making performance of users.

To support the offline measure task, following this year's assessment, we will recruit test subjects to perform a decision making task using a selection of this year runs. We anticipate that our methods will be similar to those used by Pogacar et al. [16] and Jimmy et al. [10], where test subjects are given a fixed results list and must use it to help them make a decision. The fixed results list will be defined from the runs submitted to Task 1 (retrieval task) this year. Specifically, the task will be to predict the decision the user will make at the end of the search process given a query, document ranking (results list), and relevance judgments. The user will need to decide whether a treatment is helpful for a given health issue. We will evaluate groups' performance on this task based on their prediction quality. In effect, we will perform a meta-evaluation (an evaluation of evaluation methods). We anticipate that evaluation methods may use ratios of correct vs incorrect decisions made having examined the systems' result lists, the effort (amount of interaction) required by users to reach the decisions, and other inputs to make their predictions. These avenues of evaluation are currently being explored, e.g., see van der Vegt et al. [19].

**Acknowledgement**
Thanks to Fuat Beylunioğlu and Lucas Chaves Lima for their help. This work was supported in part by AMAOS (Advanced Machine Learning for Automatic Omni-Channel Support), funded by Innovationsfonden, Denmark, and in part by the Natural Sciences and Engineering Research Council of Canada (Grant GPIN-2014-03642). Guido Zuccon is the recipient of an Australian Research Council DECRA Research Fellowship (DE180101579) and a Google Faculty Award; both partially supported this work.

# References

[1] M. Abualsaud, N. Ghelani, H. Zhang, M. D. Smucker, G. V. Cormack, and M. R. Grossman. A system for efficient high-recall retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1317–1320, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210176. URL https://doi.org/10.1145/3209978.3210176.

[2] M. Abualsaud, F. C. Beylunioğlu, M. D. Smucker, and P. R. Duimering. UWaterlooMDS at the TREC 2019 Decision Track. In *TREC*, 2019.

[3] L. Bero and D. Rennie. The Cochrane Collaboration: Preparing, Maintaining, and Disseminating Systematic Reviews of the Effects of Health Care. *JAMA*, 274(24):1935–1938, 12 1995. ISSN 0098-7484. doi: 10.1001/jama.1995.03530240045039. URL https://doi.org/10.1001/jama.1995.03530240045039.

[4] A. Bondarenko, V. Kasturia, M. Fröbe, M. Völske, B. Stein, and M. Hagen. Webis at TREC 2019: Decision Track. In *TREC*, 2019.

[5] C. E. Buckley and E. M. Voorhees. Retrieval System Evaluation. In *Experiment and Evaluation in Information Retrieval*, TREC, pages 53–78. MIT Press, 2005.

[6] W. Cui, Y. Jiang, S. Tao, and G. Hanzhang. ICTNET at Trec 2019 Decision Track. In *TREC*, 2019.

[7] J. S. Culpepper, F. Diaz, and M. D. Smucker. Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum*, 52(1): 34–90, 2018. ISSN 0163-5840. doi: 10.1145/3274784.3274788. URL `http://doi.acm.org/10.1145/3274784.3274788`.

[8] J. P. T. Higgins and S. Green. *Cochrane Handbook for Systematic Reviews of Interventions*, volume 4. John Wiley & Sons, 2011.

[9] K. Järvelin and J. Kekäläinen. Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. ISSN 1046-8188. doi: 10.1145/582415.582418. URL `http://doi.acm.org/10.1145/582415.582418`.

[10] . Jimmy, G. Zuccon, B. Koopman, and G. Demartini. Health cards for consumer health search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 35–44. ACM, 2019.

[11] Jimmy and G. Zuccon. UQ IElab at TREC 2019 Decision Track. In *TREC*, 2019.

[12] Jimmy, G. Zuccon, J. R. M. Palotti, L. Goeuriot, and L. Kelly. Overview of the clef 2018 consumer health search task. In *CLEF*, 2018.

[13] 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. ISBN 978-1-4503-4490-6. doi: 10.1145/3121050.3121072. URL `http://doi.acm.org/10.1145/3121050.3121072`.

[14] J. Palotti, G. Zuccon, P. P. Jimmy, M. Lupu, L. Goeuriot, L. Kelly, and A. Hanbury. Clef 2017 task overview: the ir task at the ehealth evaluation lab. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*, pages 1–10, 2017.

[15] 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.

[16] F. A. Pogacar, A. Ghenai, M. D. Smucker, and C. L. A. Clarke. The Positive and Negative Influence of Search Results on People's Decisions About the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '17, pages 209–216, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4490-6. doi: 10.1145/3121050.3121074. URL `http://doi.acm.org/10.1145/3121050.3121074`.

[17] M. Sinha, S. Mannarswamy, and S. Roy. Chis@ fire: Overview of the shared task on consumer health information search. In *FIRE (Working Notes)*, pages 193–196, 2016.

[18] L. Soldaini, A. Yates, E. Yom-Tov, O. Frieder, and N. Goharian. Enhancing web search in the medical domain via query clarification. *Information Retrieval Journal*, 19(1-2):149–173, 2016.

[19] A. van der Vegt, G. Zuccon, B. Koopman, and P. Bruza. A Task Completion Framework to Support Single-Interaction IR Research. *Journal of Documentation*, 74(2):289–308, 2018. doi: 10.1108/JD-09-2017-0128. URL `https://doi.org/10.1108/JD-09-2017-0128`.

[20] A. van der Vegt, G. Zuccon, B. Koopman, and A. Deacon. Impact of a search engine on clinical decisions under time and system effectiveness constraints: Research protocol. *JMIR research protocols*, 8(5):e12803, 2019.

[21] E. M. Voorhees. Evaluation by Highly Relevant Documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 74–82, New York, NY, USA, 2001. ACM. ISBN 1-58113-331-6. doi: 10.1145/383952.383963. URL `http://doi.acm.org/10.1145/383952.383963`.

[22] J. I. Westbrook, E. W. Coiera, and A. S. Gosling. Do online information retrieval systems help experienced clinicians answer clinical questions? *Journal of the American Medical Informatics Association*, 12(3):315–321, 2005.

[23] R. W. White and A. Hassan. Content bias in online health search. *ACM Trans. Web*, 8(4):25:1–25:33, Nov. 2014. ISSN 1559-1131. doi: 10.1145/2663355. URL `http://doi.acm.org/10.1145/2663355`.

[24] R. W. White and E. Horvitz. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. *ACM Transactions on Information Systems (TOIS)*, 27(4):23:1–23:37, 2009. ISSN 1046-8188. doi: 10.1145/1629096.1629101. URL `http://doi.acm.org/10.1145/1629096.1629101`.

[25] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594. ACM, 2008.