# CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving Information About Medical Symptoms

João Palotti[1], Guido Zuccon[2], Lorraine Goeuriot[3], Liadh Kelly[4], Allan Hanbury[1], Gareth J.F. Jones[5], Mihai Lupu[1], and Pavel Pecina[6⋆]

[1] Vienna University of Technology, Austria, palotti,hanbury,lupu@ifs.tuwien.ac.at
[2] Queensland University of Technology, Australia, g.zuccon@qut.edu.au
[3] Université Grenoble Alpes, France, lorraine.goeuriot@imag.fr
[4] Trinity College, Dublin, Ireland Liadh.Kelly@tcd.ie,
[5] Dublin City University, Ireland, gareth.jones@computing.dcu.ie
[6] Charles University in Prague, Czech Republic, pecina@ufal.mff.cuni.cz

**Abstract.** This paper details methods, results and analysis of the CLEF 2015 eHealth Evaluation Lab, Task 2. This task investigates the effectiveness of web search engines in providing access to medical information with the aim of fostering advances in the development of these technologies.
The problem considered in this year's task was to retrieve web pages to support information needs of health consumers that are confronted with a sign, symptom or condition and that seek information through a search engine, with the aim to understand which condition they may have. As part of this evaluation exercise, 66 query topics were created by potential users based on images and videos of conditions. Topics were first created in English and then translated into a number of other languages. A total of 97 runs by 12 different teams were submitted for the English query topics; one team submitted 70 runs for the multilingual topics.

**Key words:** Medical Information Retrieval, Health Information Seeking and Retrieval

## 1 Introduction

This document reports on the CLEF 2015 eHealth Evaluation Lab, Task 2. The task investigated the problem of retrieving web pages to support information needs of health consumers (including their next-of-kin) that are confronted with a sign, symptom or condition and that use a search engine to seek understanding about which condition they may have. Task 2 has been developed within the CLEF 2015 eHealth Evaluation Lab, which aims to foster the development

⋆ In alphabetical order, JP, GZ led Task 2; LG, LK, AH, ML & PP were on the Task 2 organising committee.

of approaches to support patients, their next-of-kin, and clinical staff in understanding, accessing and authoring health information [1].

The use of the Web as source of health-related information is a wide-spread phenomena. Search engines are commonly used as a means to access health information available online [2]. Previous iterations of this task (i.e. the 2013 and 2014 CLEFeHealth Lab Task 3 [3, 4]) aimed at evaluating the effectiveness of search engines to support people when searching for information about their conditions, e.g. to answer queries like "thrombocytopenia treatment corticosteroids length". These past two evaluation exercises have provided valuable resources and an evaluation framework for developing and testing new and existing techniques. The fundamental contribution of these tasks to the improvement of search engine technology aimed at answering this type of health information need is demonstrated by the improvements in retrieval effectiveness provided by the best 2014 system [5] over the best 2013 system [6] (using different, but comparable, topic sets).

Searching for self-diagnosis information is another important type of health information seeking activity [2]; this seeking activity has not been considered in the previous CLEF eHealth tasks, nor in other information retrieval evaluation campaigns. These information needs often arise before attending a medical professional (or to help the decision of attending). Previous research has shown that exposing people with no or scarce medical knowledge to complex medical language may lead to erroneous self-diagnosis and self-treatment and that access to medical information on the Web can lead to the escalation of concerns about common symptoms (e.g., cyberchondria) [7, 8]. Research has also shown that current commercial search engines are yet far from being effective in answering such queries [9]. This type of query is the subject of investigation in this CLEF 2015 eHealth Lab Task 2. We expected these queries to pose a new challenge to the participating teams; a challenge that, if solved, would lead to significant contributions towards improving how current commercial search engines answer health queries.

The remainder of this paper is structured as follows: Section 2 details the task, the document collection, topics, baselines, pooling strategy, and evaluation metrics; Section 3 presents the participants' approaches, while Section 4 presents their results; Section 5 concludes the paper.

## 2 The CLEF 2015 eHealth Task 2

### 2.1 The Task

The goal of the task is to design systems which improve health search, especially in the case of search for self-diagnosis information. The dataset provided to participants is comprised of a document collection, topics in various languages, and the corresponding relevance information. The collection was provided to participants after signing an agreement, through the PhysioNet website[7].

[7] `http://physionet.org/`

Participating teams were asked to submit up to ten runs for the English queries, and an additional ten runs for each of the multilingual query languages. Teams were required to number runs such as that run 1 was a baseline run for the team; other runs were numbered from 2 to 10, with lower numbers indicating higher priority for selection of documents to contribute to the assessment pool (i.e. run 2 was considered of higher priority than run 3).

### 2.2 Document Collection

The document collection provided in the CLEF 2014 eHealth Lab Task 3 [4] is also adopted in this year's task. Documents in this collection have been obtained through a large crawl of health resources on the Web; the collection contains approximately one million documents and originated from the Khresmoi project[8] [10]. The crawled domains were predominantly health and medicine sites, which were certified by the HON Foundation as adhering to the HONcode principles (appr. 60–70% of the collection), as well as other commonly used health and medicine sites such as Drugbank, Diagnosia and Trip Answers[9]. Documents consisted of web pages on a broad range of health topics and were likely targeted at both the general public and healthcare professionals. They were made available for download in their raw HTML format along with their URLs to registered participants.

### 2.3 Topics

Queries were manually built with the following process: images and videos related to medical symptoms were shown to users, who were then asked which queries they would issue to a web search engine if they, or their next-of-kin, were exhibiting such symptoms. Thus, these queries aimed to simulate the situation of health consumers seeking information to understand symptoms or conditions they may be affected by; this is achieved using imaginary or video stimuli. This methodology for eliciting circumlocutory, self-diagnosis queries was shown to be effective by Stanton et al. [11]; Zuccon et al. [9] showed that current commercial search engines are yet far from being effective in answering such queries.

Following the methodology in [9, 11], 23 symptoms or conditions that manifest with visual or audible signs (e.g. ringworm or croup) were selected to be presented to users to collect queries. A cohort of 12 volunteer university students and researchers based in the organisers' institutions generated the queries. English was the mother-tongue for all volunteers and they had no particular prior knowledge about the symptoms or conditions, nor they had any specific medical background: this cohort was then somehow representative of the average user of web search engines seeking health advice (although they had a higher

[8] Medical Information Analysis and Retrieval, `http://www.khresmoi.eu`

[9] Health on the Net, `http://www.healthonnet.org`, `http://www.hon.ch/HONcode/Patients-Conduct.html`, `http://www.drugbank.ca`, `http://www.diagnosia.com`, and `http://www.tripanswers.org`

education level than the average level). Each volunteer was given 10 conditions for which they were asked to generate up to 3 queries per condition (thus each condition/image pair was presented to more than one assessor[10]). An example of images and instructions provided to the volunteers is given in Figure 1[11].

```
Imagine you are experiencing the health problem shown below.
Please provide 3 search queries that you would issue to find out what is wrong.
Instructions:
* You must provide 3 distinct search queries.
* The search queries must relate to what you see below.
```

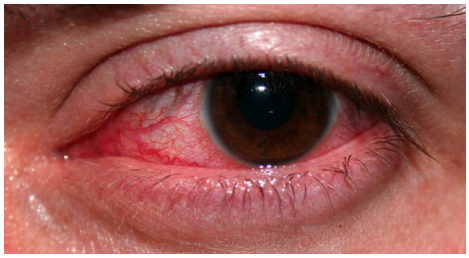

**Fig. 1.** An example of instructions and images provided to volunteers for generating potential search queries.

A total of 266 possible unique queries were collected; of these, 67 queries (22 conditions with 3 queries and 1 condition with 1 query) were selected to be used in this year's task. Queries were selected by randomly picking one query per condition (we called this the *pivot* query), and then manually selecting the query that appeared most similar (called *most*) and the one that appeared least similar (called *least*) to the pivot query. Candidates for the *most* and *least* queries were identified independently by three organisers and then majority voting was used to establish which queries should be selected. This set of queries formed the *English query set* distributed to participants to collect runs.

In addition, we developed translations of this query set into Arabic (AR), Czech (CS), German (DE), Farsi (FA), French (FR), Italian (IT) and Portuguese (PT); these formed the *multilingual query sets* which were made available to participants for submission of multilingual runs. Queries were translated by medical experts available at the organisers institutions.

After the query set was released, numbered *qtest*1-*qtest*67, one typo was found in query *qtest*62, which could compromise the translations. In order to keep consistency between the English query and all translations made by the experts, *qtest*62 was excluded. Thus, the *final query set* used in the CLEF 2015

[10] With exception of one condition, for which only one query could be generated.

[11] Note that additional instructions were given to volunteers at the start and end of the task, including training and de-briefing.

eHealth Lab Task 2 for both English and multilingual queries consisted of 66 queries.

An example of one of the query topics generated from the image shown in Figure 1 is provided in Figure 2. To develop their submissions, participants were only given the `query` field of each query topic, that is, teams were unaware of the query type (pivot, most, least), the target condition and the image or video that was shown to assessors to collect queries.

```
<topics>
...
<top>
<num>qtest.23</num>
<query>red bloodshot eyes</query>
<disease>non-ulcerative sterile keratitis</disease>
<type>most</type>
<query_index>22</query_index>
</top>
...
</topics>
```

**Fig. 2.** Example query topic generated from the image of Figure 1. This query is of type *most* and refers to the image condition 22 (as indicated by the field `query_index`).

### 2.4 Relevance Assessment

Relevance assessments were collected by pooling participants' submitted runs as well as baseline runs (see below for a description of pooling methodology and baseline runs). Assessment was performed by five paid medical students employed at the Medizinische Universität Graz (Austria); assessors used Relevation! [12] to visualise and judge documents. For each document, assessors had access to the query the document was retrieved for, as well as the target symptom or condition that was used to obtained the query during the query generation phase.

Target symptoms or conditions were used to provide the relevance criteria assessors should judge against; for example for query *qtest*1 – "many red marks on legs after traveling from US" (the condition used for generating the query was "Rocky Mountain spotted fever (RMSF)"), the relevance criterion read "Relevant documents should contain information allowing the user to understand that they have Rocky Mountain spotted fever (RMSF).". Relevance assessments were provided on a three point scale: 0, Not Relevant; 1, Somewhat Relevant; 2, Highly Relevant.

Along with relevance assessments, readability judgements were also collected for the assessment pool. The notion of readability and understandability of information is of important concern when retrieving information for health consumers [13]. It has been shown that if the readability of information is accounted for in the evaluation framework, judgements of relative system effectiveness can vary with respect to taking into account (topical) relevance only [14] (this was the case also when considering the CLEF 2013 and 2014 eHealth Evaluation Labs).

Readability assessments were collected by asking the assessors whether they believed a patient would understand the retrieved document. Assessments were provided on a four point scale: 0, "It is very technical and difficult to read and understand"; 1, "It is somewhat technical and difficult to read and understand"; 2, "It is somewhat easy to read and understand"; 3, "It is very easy to read and understand".

### 2.5 Example Topics

A different set of 5 queries was released to participants as example queries (called *training*) to help develop their systems (both in English and the other considered languages). These queries were released together with associated relevance assessments, obtained by evaluating a pool of 112 documents retrieved by a set of baseline retrieval systems (TF-IDF, BM25, Language Model with Dirichlet smoothing as implemented in Terrier [15], with the associated default parameter values); the pool was formed by sampling the top 10 retrieved documents for each query. Note that, given the very limited pool and system sample sizes, these example queries should not be used to evaluate, tune or train systems.

### 2.6 Baseline Systems

The organisers generated baseline runs using BM25, TF-IDF and Language Model with Dirichlet smoothing, as well as a set of benchmark systems that ranked documents by estimating both (topical) relevance and readability. Table 1 shows the 13 baseline systems created, 7 of them took into consideration some estimation of text readability. No baselines were created for the multilingual queries.

The first 6 baselines, named *baseline1-6*, were created using either Xapian or Lucene as retrieval toolkit. We vary the retrieval model used, including BM25 (with parameters $k_1 = 1, k_2 = 0, k_3 = 1$ and $b = 0.5$) in *baseline1*, Vector Space Model (VSM) with TF-IDF weighting (the default Lucene implementation) in *baseline2*, and Language Model (LM), with Dirichlet smoothing with $\mu = 2,000$ in *baseline3*. Our preliminary runs based on the 2014 topics showed that removing HTML tags from documents in this collection could lead to higher retrieval effectiveness when using BM25 and LM. We used the python package BeautifulSoap (BS4)[12] to parse the HTML files and remove HTML tags. Note that it

[12] https://pypi.python.org/pypi/beautifulsoup4

does not remove the boilerplate from the HTML (such as headers or navigation menus), being one of the simplest approaches to clean a HTML page and prepare it to serve as the input of readability formulas [16] (see below). Baselines 4, 5 and 6 implement the same methods as in baselines 1, 2 and 3, respectively, but execute a query that has been enhanced by augmenting the original query with the known target disease names. Note that the target disease names were only known to the organisers, participants had no access to this information.

For the baseline runs that take into account readability estimations, we used two well-known automatic readability measures: the Dale-Chall measure [17] and Flesch-Kincaid readability index [18]. The python package ReadabilityCalculator[13] was used to compute the readability measures from the cleansed web documents. We also tested a readability measure based on the frequency of words in a large collection such as Wikipedia; the intuition behind this measure is that an easy text would contain a large number of common words with high frequency in Wikipedia, while a technical and difficult text would have a large number of rare words, characterised by a low frequency in Wikipedia. In order to retrieving documents accounting for their readability levels, we first generate a readability score $Read(d)$ for each document $d$ in the collection using one of the three measures above. We then combine the readability score of a document with its relevance score $Rel(d)$ generated by some retrieval model. Three score combination methods were considered:

1. Linear combination: $Score(d) = \alpha \times Rel(d) + (1.0 - \alpha) \times Read(d)$, where $\alpha$ is a hyperparameter and $0 \leq \alpha \leq 1$ (in $readability1$ $\alpha$ is 0.9)
2. Direct Multiplication: $Score(d) = Rel(doc) \times Read(d)$
3. Inverse Logarithm: $Score(d) = \frac{Rel(doc)}{\log(Read(d))}$

Table 1 shows the settings of retrieval model, HTML processing, readability measure and query expansion or score combination method that were considered to produce the 7 readability baselines used in the task.

### 2.7 Pooling Methodology

In Task 2, for each query, the top 10 documents returned in runs 1, 2 and 3 produced by the participants[14] were pooled to form the relevance assessment pool. In addition, the baseline runs developed by the organisers were also pooled with the same methodology used for participants runs. A pool depth of 10 documents was chosen because this task resembles web-based search, where often users consider only the first page of results (that is, the first 10 results). Thus, this pooling methodology allowed a full evaluation of the top 10 results for the 3 submissions with top priority for each participating team. The pooling of more submissions or a deeper pool, although preferable, was ruled out because of the limited availability of resources for document relevance assessment.

[13] https://pypi.python.org/pypi/ReadabilityCalculator/

[14] With the exclusion of multilingual submissions, for which runs were not pooled due to the larger assessment effort pooling these runs would have required. Note that only one team submitted multilingual runs.

**Table 1.** Scheme showing the settings of retrieval model, HTML processing, readability measure and query expansion or score combination used to generate the organisers baselines.

| System | Index | Model | Cleaning | Expansion/Combination | Readability |
|---|---|---|---|---|---|
| **baseline1** | Xapian | BM25 | BS4 | - | - |
| **baseline2** | Lucene | VSM | - | - | - |
| **baseline3** | Lucene | LM | BS4 | - | - |
| **baseline4** | Xapian | BM25 | BS4 | Disease Name added | - |
| **baseline5** | Lucene | VSM | - | Disease Name added | - |
| **baseline6** | Lucene | LM | BS4 | Disease Name added | - |
| **readability1** | Xapian | BM25 | BS4 | Linear Combination | Dale-Chall |
| **readability2** | Xapian | BM25 | BS4 | Direct Multiplication | Wikipedia Frequency |
| **readability3** | Xapian | BM25 | BS4 | Inverse Logarithm | Dale-Chall |
| **readability4** | Xapian | BM25 | BS4 | Inverse Logarithm | Flesch-Kincaid |
| **readability5** | Lucene | VSM | - | Direct Multiplication | Wikipedia Frequency |
| **readability6** | Lucene | VSM | BS4 | Inverse Logarithm | Dale-Chall |
| **readability7** | Lucene | VSM | BS4 | Inverse Logarithm | Flesch-Kincaid |

### 2.8 Multilingual Evaluation: Additional Pooling and Relevance Assessments

Because only one team submitted runs for the multilingual queries and only limited relevance assessment capabilities were available through the paid medical students that performed the assessment of submissions for the English queries, multilingual runs were not considered when forming the pools for relevance assessments. However, additional relevance assessments were sought through the team that participated in the multilingual task: they were thus asked to perform a self-assessment of the submissions they produced. A new pool of documents was sampled with the same pooling methodology used for English runs (see the previous section). Documents that were already judged by the official assessors were excluded from the pool with the aim to limit the additional relevance assessment effort required by the team.

Additional relevance assessments for the multilingual runs were then performed by a medical doctor (native Czech speaker with fluent English) associated with Team CUNI [22]. The assessor was provided with the same instructions and assessment system that the official assessors used. Assessments were collected and aggregated with those provided by the official relevance assessors to form the multilingual *merged* qrels. These qrels should be used with caution: at the moment of writing this paper, it is unknown whether these multilingual assessments are comparable with those compiled by the original, also medically trained, assessors. The collection of further assessments from the team to verify their agreement with the official assessors is left for future work. Another limitation of these additional relevance assessments is that only one system that considered multilingual queries, that developed by team CUNI, was sampled and thus it may further bias the assessment of retrieval systems with respect to how multilingual queries are coped with.

### 2.9 Evaluation Metrics

Evaluation was performed in terms of graded and binary assessments. Binary assessments were formed by transforming the graded assessments such that label 0 was maintained (i.e. irrelevant) and labels 1 and 2 were converted to 1 (relevant). Binary assessments for the readability measures were obtained similarly, with labels 0 and 1 being converted into 0 (not readable) and labels 2 and 3 being converted into 1 (readable).

System evaluation was conducted using precision at 10 (p@10) and normalised discounted cumulative gain [19] at 10 (nDCG@10) as the primary and secondary measures, respectively. Precision was computed using the binary relevance assessments; nDCG was computed using the graded relevance assessments. These evaluation metrics were computed using `trec_eval` with the following commands:

```
./trec_eval -c -M1000 qrels.clef2015.test.bin.txt runName
./trec_eval -c -M1000 -m ndcg_cut qrels.clef2015.test.graded.txt runName
```

respectively to compute precision and nDCG values.

A separate evaluation was conducted using both relevance assessments and readability assessments following the methods in [14]. For all runs, Rank Biased Precision (RBP) [20] was computed along with readability-biased modifications of RBP, namely uRBP (using the binary readability assessments) and uRBPgr (using the graded readability assessments).

The RBP parameter $\rho$ which attempts to model user behaviour[15] (RBP persistence parameter) was set to 0.8 for all variations of this measure, following the findings of Park and Zhang [21].

To compute uRBP, readability assessments were mapped to binary classes, with assessments 0 and 1 (indicating low readability) mapped to value 0 and assessments 2 and 3 (indicating high readability) mapped to value 1. Then, uRBP (up to rank $K$) was calculated according to

$$uRBP = (1-\rho)\sum_{k=1}^{K}\rho^{k-1}r(k)u(k) \tag{1}$$

where $r(k)$ is the standard RBP gain function that is 1 if the document at rank $k$ is relevant and 0 otherwise; $u(k)$ is a similar gain function but for the readability dimension and is 1 if the document at $k$ is readable (binary class 1), and zero otherwise (binary class 0).

To compute uRBPgr, i.e. the graded version of uRBP, each readability label was mapped to a different gain value. Specifically, label 0 was assigned gain 0 (least readability, no gain), label 1 gain 0.4, label 2 gain 0.8 and label 3 gain 1 (highest readability, full gain). Thus, a document that is somewhat difficult to read does still generate a gain, which is half the gain generated by a document

[15] High values of $\rho$ representing persistent users, low values representing impatient users.

that is somewhat easy to read. These gains are then used to evaluate the function $u(k)$ in Equation 1 to obtain uRBPgr.

The readability-biased evaluation was performed using `ubire`[16], which is publicly available for download.

## 3 Participants and Approaches

### 3.1 Participants

This year, 52 groups registered for the task on the web site, 27 got access to the data and 12 submitted any run for task 2. The groups are from 9 countries in 4 continents as listed in Table 2. 7 out of the 12 participants had never participated in this task before.

**Table 2.** Participants for task 2 and their total number of submissions.

| Continent | Country | Team Name | Runs Submitted | |
|---|---|---|---|---|
| | | | English | Multilingual |
| Africa | Botswana | UBML | 10 | - |
| | Tunisia | Miracl | 5 | - |
| America | Canada | GRIUM | 7 | - |
| | Canada | YORKU | 10 | - |
| Asia | China | ECNU | 10 | - |
| | China | FDUSGInfo | 10 | - |
| | China | USST | 10 | - |
| | South Korea | KISTI | 8 | - |
| | Thailand | KU-CS | 4 | - |
| | Vietnam | HCMUS | 8 | - |
| Europe | Czech Republic | CUNI | 10 | 70 |
| | France | LIMSI | 5 | - |
| **Total** | **9 Countries** | **12 Teams** | **97** | **70** |

### 3.2 Participant Approaches

*Team CUNI* [22] used the Terrier toolkit to produce their submissions. Runs explored three different retrieval models: Bayesian smoothing with Dirichlet prior, Per-field normalisation (PL2F) and LGD. Query expansion using the UMLS metathesaurus was explored by considering terms assigned to the same concept as synonymous and choosing the terms with the highest inverse document-frequency. Blind relevance feedback was also used as contrasting technique. Finally, they also experimented with linear interpolations of the search results produced by the above techniques.

[16] `https://github.com/ielab/ubire`

*Team ECNU* [23] explored query expansion and learning to rank. For query expansion, Google was queried and the titles and snippets associated with the top ten web results were selected. Medical terms were then extracted from these resources by matching them with terms contained in MeSH; the query was then expanded using those medical terms that appeared more often than a threshold. As Learning to Rank approach, Team ECNU combined scores and ranks from BM25, PL2 and BB2 into a six-dimensional vector. The 2013 and 2014 CLEF eHealth tasks were used to train the system and a Random Forest classifier was use to calculate the new scores.

*Team FDUSGInfo* explored query expansion methods that use a range of knowledge resources to improve the effectiveness of a statistical Language Model baseline. The knowledge sources that have been considering for drawing expansion terms are MeSH and Freebase. Different techniques were evaluated to select the expansion terms, including manual term selection. Team FDUSGInfo, unfortunately, did not submit their working notes and thus the details of their methods are unknown.

*Team GRIUM* [25] explored the use of concept based query expansion. Their query expansion mechanism exploited Wikipedia articles and UMLS Concept definitions and were compared to a baseline method based on Dirichlet smoothing.

*Team KISTI* [26] focused on re-ranking approaches. Lucene was used for indexing and initial search, and the baseline used the query likelihood model with Dirichlet smoothing. They explored three approaches for re-ranking: explicit semantic analysis (ESA), concept-based document centrality (CBDC), and cluster-based external expansion model (CBEEM). Their submissions evaluated these re-ranking approaches as well as their combinations.

*Team KUCS* [27] implemented an adaptive query expansion. Based on the results returned by a query performance prediction approach, their method selected the query expansion that is hypothesised to be the most suitable for improving effectiveness. An additional process was responsible for re-ranking results based on readability estimations.

*Team LIMSI* [28] explored query expansion approaches that exploit external resources. Their first approach used MetaMap to identify UMLS concepts from which to extract medical terms to expand the original queries. Their second approach used a selected number of Wikipedia articles describing the most common diseases and conditions along with a selection of MedlinePlus; for each query the most relevant articles from these corpora are retrieved and their titles used to expand the original queries, which are in turn used to retrieve relevant documents from the task collection.

*Team Miracl* [29]'s submissions were based on blind relevance feedback combined with term selection using their previous work on modeling semantic relations between words. Their baseline run was based on the traditional Vector Space Model and the Terrier toolkit. The other runs employed the investigated method by varying settings of two method parameters: the first controlling the number of highly ranked documents from the initial retrieval step and the second controlling the degree of semantic relationship of the expansion terms.

*Team HCMUS* [30] experimented with two approaches. The first was based on concept-based retrieval where only medical terminological expressions in documents were retained, while other words were filtered-out. The second was based on query expansion with blind relevance feedback. Common to all their approaches was the use of Apache Lucene and a bag-of-word baseline based on Language Modelling with Dirichlet smoothing and standard stemming and stopword removal.

*Team UBML* [31] investigated the empirical merits of query expansion based on KL divergence and the Bose-Einstein 1 model for improving a BM25 baseline. The query expansion process selected terms from the local collection or two external collections. Learning to rank was also investigated along a Markov Random Fields approach.

*Team USST* [32] used BM25 as a baseline system and explored query expansion approaches. They investigated pseudo relevance feedback approaches based on Kullback-Liebler Divergence and Bose-Einstein models.

*Team YorkU* [33] explored BM25 and Divergence from Randomness methods as provided by the Terrier toolkit, along with the associated relevance feedback retrieval approaches.

## 4 Results and Findings

### 4.1 Pooling and Coverage of Relevance Assessments

A total of 8,713 documents were assessed. Of these, 6,741 (77.4%) were assessed as irrelevant (0), 1,515 (17.4%) as somewhat relevant (1), 457 (5.2%) as highly relevant (2). For readability assessments, the recorded distribution was: 1,145 (13.1%) documents assessed as difficult (0), 1,568 (18.0%) as somewhat difficult (1), 2,769 (31.8%) as somewhat easy (2), and 3,231 (37.1%) as easy (3).

Table 3 details the coverage of the relevance assessments with respect to the participant submissions, averaged over the whole query set. While in theory all runs 1-3 should have full coverage (100%), in practice a small portion of documents included in the relevance assessment pool were left unjudged because the documents were not in the collection (participants provided an invalid document identifier) or the page failed to render in the relevance assessment toolkit (for example because the page contained redirect scripts or other scripts that were

**Table 3.** Coverage of the relevance assessments for the top 10 results submitted by participants in the task: 100% means that all top 10 results for all queries have been assessed; 90% means that, on average, 9 out of 10 documents in the top 10 results have been assessed, with one document being left unjudged.

| Run | Baseline | Readab. | CUNI | ENUC | FDUSG. | GRIUM | KISTI | KUCS | LIMSI | Miracl | HCMUS | UBML | USST | YorkU | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 99.98 | 100.0 | 100.0 | 99.98 | 98.77 | 100.0 | 100.0 | 99.64 | 99.83 | 99.98 | 99.92 | 99.92 | 100.0 | 99.62 | 99.83 |
| **2** | 99.82 | 99.98 | 100.0 | 99.88 | 98.77 | 100.0 | 99.98 | 98.77 | 99.92 | 99.98 | 99.89 | 100.0 | 100.0 | 100.0 | 99.79 |
| **3** | 99.98 | 99.95 | 99.94 | 99.95 | 98.77 | 100.0 | 100.0 | 92.61 | 99.85 | 100.0 | 99.79 | 100.0 | 100.0 | 100.0 | 99.35 |
| **4** | 93.64 | 94.65 | 99.95 | 99.86 | 98.08 | 99.65 | 99.80 | 91.58 | 92.00 | 96.82 | 97.65 | 98.38 | 98.64 | 99.98 | 97.19 |
| **5** | 92.61 | 99.15 | 99.58 | 96.00 | 97.91 | 99.94 | 99.58 | - | 92.00 | 99.15 | 94.67 | 98.42 | 98.30 | 99.85 | 97.47 |
| **6** | 93.74 | 98.89 | 99.23 | 98.11 | 91.65 | 99.98 | 99.73 | - | - | - | 93.12 | 98.58 | 97.91 | 99.68 | 97.33 |
| **7** | - | 97.33 | 99.79 | 96.56 | 91.65 | 99.98 | 99.70 | - | - | - | 94.65 | 99.48 | 96.24 | 99.53 | 97.49 |
| **8** | - | - | 99.98 | 98.76 | 91.65 | - | 99.73 | - | - | - | 93.14 | 98.29 | 95.85 | 99.23 | 97.08 |
| **9** | - | - | 99.61 | 99.79 | 91.33 | - | - | - | - | - | - | 98.45 | 95.24 | 98.83 | 97.21 |
| **10** | - | - | 97.94 | 98.70 | 91.33 | - | - | - | - | - | - | 97.70 | 95.06 | 98.33 | 96.51 |
| **Mean** | 96.63 | 98.57 | 99.60 | 98.76 | 94.99 | 99.94 | 99.81 | 95.65 | 96.72 | 99.19 | 96.60 | 98.92 | 97.72 | 99.51 | 98 |

not executed within Relevation[17]). Overall, the mean coverage for runs 1-3 was above 99%, with only run 3 from team KUCS being sensibly below this value. This suggests that the retrieval effectiveness for runs 1-3 can be reliably measured. The coverage beyond submissions 3 is lower but always above 90% (and the mean above 95%); this suggest that the evaluation of runs 4-10 in terms of precision at 10 may be underestimated of an average maximum of 0.05 points over the whole query set.

Table 4 details the coverage of relevance assessment for the multilingual runs. As mentioned in Section 2.8, due to limited relevance assessment availability, only the English runs were considered when forming the pool for relevance assessment. The coverage of these relevance assessments with respect to the top 10 documents ranked by each participants' submissions is shown in the columns marked as *Eng.* in Table 4. An additional document pool, made using only documents in *runs1-3* of multilingual submissions, was created to further increase the coverage of multilingual submissions; the coverage of the union of the original assessments and these additional ones (referred to as *merged*) is shown in the columns marked as *Merged* in Table 4 for the multilingual runs. The *merged* set of relevance assessments was enough to provide a fairly high coverage for all runs, including those not in the pool (i.e., runs beyond number 3), with a minimal coverage of 97%; this is likely because only one team submitted runs for the multilingual challenge, thus providing only minimal variation in terms of top retrieval results.

### 4.2 Evaluation Results and Findings

Table 5 reports the evaluation of the participants submissions and the organisers baselines based on P@10 and nDCG@10 for English queries. The evaluation based on RBP and the readability measures is reported in Table 6.

Most of the approaches developed by team ECNU obtain significantly higher values of P@10 and nDCG@10 compared to the other participants, demonstrat-

[17] Note that before the relevance assessment exercise started, we removed the majority of scripts from the pooled pages to avoid this problem.

**Table 4.** Coverage of the relevance assessments for the top 10 results submitted by CUNI in the multilingual evaluation. As described in Section 2.8, two set of qrels were used: those for the English task (*Eng.*), and those produced by merging the assessments for English queries and the ones for multilingual queries (*Merged*.)

| Run | AR | | CS | | DE | | FA | | FR | | IT | | PT | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Eng. | Merged | Eng. | Merged | Eng. | Merged | Eng. | Merged | Eng. | Merged | Eng. | Merged | Eng. | Merged |
| **1** | 95.32 | 99.97 | 94.52 | 99.94 | 95.21 | 99.80 | 95.59 | 99.91 | 95.14 | 99.91 | 95.48 | 99.95 | 95.76 | 99.94 |
| **2** | 94.95 | 99.91 | 93.88 | 99.82 | 94.85 | 99.82 | 95.36 | 99.89 | 94.59 | 99.92 | 95.35 | 99.85 | 95.56 | 99.91 |
| **3** | 94.64 | 99.91 | 93.74 | 99.86 | 94.70 | 99.91 | 95.11 | 99.91 | 94.65 | 99.92 | 94.89 | 99.89 | 95.18 | 99.83 |
| **4** | 95.20 | 99.77 | 94.09 | 99.79 | 95.11 | 99.77 | 95.62 | 99.79 | 94.88 | 99.88 | 95.58 | 99.89 | 95.83 | 99.85 |
| **5** | 95.00 | 98.03 | 94.02 | 97.32 | 94.98 | 97.47 | 95.29 | 97.70 | 94.67 | 98.56 | 95.14 | 98.97 | 95.65 | 98.33 |
| **6** | 95.03 | 98.03 | 94.11 | 98.56 | 94.35 | 98.26 | 95.42 | 97.71 | 94.59 | 98.56 | 95.15 | 99.05 | 95.91 | 98.30 |
| **7** | 94.73 | 97.73 | 94.29 | 97.36 | 96.47 | 98.91 | 94.85 | 97.30 | 96.00 | 99.29 | 94.76 | 98.98 | 95.42 | 97.88 |
| **8** | 95.03 | 98.12 | 94.45 | 97.21 | 96.11 | 98.21 | 95.42 | 97.68 | 95.89 | 98.73 | 95.18 | 98.94 | 95.94 | 98.27 |
| **9** | 94.64 | 98.29 | 94.00 | 96.52 | 95.47 | 97.58 | 94.62 | 97.73 | 95.33 | 98.29 | 95.00 | 98.48 | 94.80 | 98.18 |
| **10** | 95.33 | 99.55 | 94.38 | 97.29 | 96.62 | 98.76 | 95.70 | 99.48 | 96.17 | 99.24 | 95.59 | 99.70 | 95.94 | 99.48 |
| **Mean** | 94.99 | 98.93 | 94.15 | 98.37 | 95.39 | 98.85 | 95.30 | 98.71 | 95.19 | 99.23 | 95.21 | 99.37 | 95.60 | 99.00 |

ing about 40% increase in effectiveness in their best run compared to the runner-up team (KISTI). The best submission developed by the organisers and based on both relevance and readability estimates has been proved difficult to outperform by most teams (only 4 out of 12 teams obtained higher effectiveness). The pooling methodology does not appear to have significantly influenced the evaluation of non-pooled submissions, as demonstrated by the fact that the best runs of some teams are not those that were fully pooled (e.g. team KISTI, team CUNI, team GRIUM).

There are no large differences between system rankings produced using P@10 or nDCG@10 as evaluation measure (Kendall $\tau = 0.88$). This is unlike when readability is also considered in the evaluation (the Kendall $\tau$ between system rankings obtained with P@10 or uRBP is 0.76). In this latter case, while ECNU's submissions are confirmed to be the most effective, there are large variations in system rankings when compared to those obtained considering relevance judgements only. In particular, runs from team KISTI, which in the relevance-based evaluation were ranked among the top 20 runs, are not performing as well when considering also readability, with their top run (KISTI_EN_RUN.7) being ranked only 37th according to uRBP.

The following considerations could be drawn when comparing the different methods employed by the participating teams. Query expansion is found to often improve results. In particular, team ECNU obtained the highest effectiveness among the systems that took part in this task; this was achieved when query expansion terms are mined from Google search results returned for the original queries (ECNU_EN_Run.3). This approach indeed obtained higher effectiveness compared to learning-to-rank alternatives (ECNU_EN_Run.10). The results of team UBML show that query expansion using the Bose-Einstein model 1 and the local collection works better than other query expansion methods and external collections. Team USST also found that query expansion was effective to improve results, however they found that the Bose-Einstein models did not

provide improvements over their baseline, while the Kullback-Liebler Divergence based query expansion provided minor improvements. Health-specific query expansion methods based on the UMLS were shown to be effective above common baselines and other considered query expansion methods by Team LIMSI and GRIUM (this form of query expansion was the only one that delivered higher effectiveness than their baseline).Team KISTI found that the combination of concept-based document centrality (CBDC) and cluster- based external expansion model (CBEEM) improved the results best. Few teams did not observe improvements over their baselines; this was the case for teams KUCS, Miracl, FDUSGInfo and HCMUS.

Tables 7 and 8 report the evaluation of the multilingual submissions based on P@10 and nDCG@10; results are reported with respect to both the original qrels (obtained by sampling English runs only) and the additional qrels (obtained by sampling also multilingual runs, but using a different set of assessors); see Section 2.8 for details about the difference between these relevance assessments. Only one team (CUNI) participated in the multilingual task; they also submitted to the English-based task and thus it is possible to discuss the effectiveness of their retrieval system when answering multilingual queries compared to that achieved when answering English queries.

The evaluation based on the original qrels allows us to compare multilingual runs directly with English runs. Note that the original relevance assessments exhibit a level of coverage for the multilingual runs that is similar to those obtained for English submissions numbered 4-10. The evaluation based on the additional qrels (merged) allows analysis of the multilingual runs using the same pooling method used for English runs; thus submissions 1-3 for the multilingual runs can be directly compared to the corresponding English ones, at the net of differences in expertise, sensibility and systematic errors between the paid medical assessors and the volunteer, student self-assessor used to gather judgements for the multilingual runs.

When only multilingual submissions are considered, it can be observed that there is not a language in which CUNI's system is more effective: e.g. submissions that considered Italian queries are among the best performing with original assessments and are the best performing with the additional assessments, but differences in effectiveness among top runs for different languages are not statistically significant. However, it can be observed that none of CUNI's submissions that addressed queries expressed in not European languages (Farsi and Arabic) are among the top ranked systems, regardless of the type of relevance assessments.

The use of the additional relevance assessments naturally translates in observing increased retrieval effectiveness across all multilingual runs (because some of the documents in the top 10 ranks that were not assessed, and thus irrelevant, in the original assessments may have been marked as relevant in the additional assessments). However, a noteworthy observation is that the majority of the most effective runs according to the additional assessments are those that were not

fully sampled to form the relevance assessment pools (i.e. runs 4-10, as opposed to the pooled runs 1-3).

When the submissions of team CUNI are compared across English and multilingual queries, it is possible to observe that the best multilingual runs do not outperform English runs (unlike when the same comparison was instructed in the 2014 task [4]), regardless of the type of relevance assessments. This result does not come as unexpected and it indicates that the translation from a foreign language to English as part of the retrieval process does degrade the quality of queries (in terms of retrieval effectiveness), suggesting that more work is needed to bridge the gap in effectiveness between English and multilingual queries when these are used to retrieve English content.

## 5 Conclusions

This paper has described methods, results and analysis of the CLEF 2015 eHealth Evaluation Lab, Task 2. The task considered the problem of retrieving web pages for people seeking health information regarding unknown conditions or symptoms. 12 teams participated in the task; the results have shown that query expansion plays an important role in improving search effectiveness. The best results were achieved by a query expansion method that mined the top results from the Google search engine. Despite the improvements over the organisers' baselines achieved by some teams, further work is needed to sensibly improve search in this context, as only about half of the top 10 results retrieved by the best system were found to be relevant.

As a by-product of this evaluation exercise, the task contributes to the research community a collection with associated assessments and evaluation framework (including readability evaluation) that can be used to evaluate the effectiveness of retrieval methods for health information seeking on the web. Queries, assessments and participants runs are publicly available at `http://github.com/CLEFeHealth/CLEFeHealth2015Task2`.

## Acknowledgement

This task has been supported in part by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°257528 (KHRESMOI), by Horizon 2020 program (H2020-ICT-2014-1) under grant agreement n° 644753 (KCONNECT), by the Austrian Science Fund (FWF) project n° I1094-N23 (MUCKE), and by the Czech Science Foundation (grant number P103/12/G084). We acknowledge the time of the people involved in the translation and relevance assessment tasks, in special we want to thank Dr. Johannes Bernhardt-Melischnig (Medizinische Universitat Graz) for coordinating the recruitment and management of the paid medical students that participated in the relevance assessment exercise.

## References


1. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J., Zuccon, G.: Overview of the clef ehealth evaluation lab 2015. In: CLEF 2015 - 6th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer (September 2015)
2. Fox, S.: Health topics: 80% of internet users look for health information online. Pew Internet & American Life Project (2011)
3. Goeuriot, L., Jones, G., Kelly, L., Leveling, J., Hanbury, A., Müller, H., Salanterä, S., Suominen, H., Zuccon, G.: ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. In: Online Working Notes of CLEF, CLEF (2013)
4. Goeuriot, L., Kelly, L., Lee, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Gareth J.F. Jones, H.M.: ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. In: CLEF 2014 Evaluation Labs and Workshop: Online Working Notes, Sheffield, UK (2014)
5. Shen, W., Nie, J.Y., Liu, X., Liui, X.: An investigation of the effectiveness of concept-based approach in medical information retrieval grium@ clef2014ehealthtask 3. Proceedings of the ShARe/CLEF eHealth Evaluation Lab (2014)
6. Zhu, D., Wu, S., James, M., Carterette, B., Liu, H.: Using discharge summaries to improve information retrieval in clinical domain. Proceedings of the ShARe/-CLEF eHealth Evaluation Lab (2013)
7. Benigeri, M., Pluye, P.: Shortcomings of health information on the internet. Health promotion international **18**(4) (2003) 381–386
8. White, R.W., Horvitz, E.: Cyberchondria: studies of the escalation of medical concerns in web search. ACM TOIS **27**(4) (2009) 23
9. Zuccon, G., Koopman, B., Palotti, J.: Diagnose this if you can: On the effectiveness of search engines in finding medical self-diagnosis information. In: Advances in Information Retrieval. Springer (2015) 562–567
10. Hanbury, A., Müller, H.: Khresmoi – multimodal multilingual medical information search. In: MIE village of the future. (2012)
11. Stanton, I., Ieong, S., Mishra, N.: Circumlocution in diagnostic medical queries. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, ACM (2014) 133–142
12. Koopman, B., Zuccon, G.: Relevation!: an open source system for information retrieval relevance assessment. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, ACM (2014) 1243–1244
13. Walsh, T.M., Volsko, T.A.: Readability assessment of internet-based consumer health information. Respiratory care **53**(10) (2008) 1310–1315
14. Zuccon, G., Koopman, B.: Integrating understandability in the evaluation of consumer health search engines. In: Medical Information Retrieval Workshop at SIGIR 2014. (2014) 32
15. Ounis, I., Lioma, C., Macdonald, C., Plachouras, V.: Research directions in terrier. Novatica UPGRADE Special Issue on Web Information Access (2007)
16. Palotti, J., Zuccon, G., Hanbury, A.: The influence of pre-processing on the estimation of readability of web documents. In: Proceedings of the 24th ACM International Conference on Conference on Information and Knowledge Management (CIKM). (2015)

17. Kincaid, J., Fishburne, R., Rogers, R., Chissom, B.: Derivation of New Readability Formulas for Navy Enlisted Personnel. Technical report (1975)
18. Kincaid, J., Fishburne, R., Rogers, R., Chissom, B.: Derivation of New Readability Formulas for Navy Enlisted Personnel. Technical report (1975)
19. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems **20**(4) (2002) 422–446
20. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions on Information Systems (TOIS) **27**(1) (2008) 2
21. Park, L.A., Zhang, Y.: On the distribution of user persistence for rank-biased precision. In: Proceedings of the 12th Australasian document computing symposium. (2007) 17–24
22. Saleh, S., Bibyna, F., Pecina, P.: CUNI at the CLEF 2015 eHealth Lab Task 2. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2015)
23. Song, Y., He, Y., Hu, Q., He, L.: ECNU at 2015 eHealth Task 2: User-centred Health Information Retrieval. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2015)
24. received, N.: Missing. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2015)
25. Liu, X.J., Nie, J.Y.: Bridging Layperson's Queries with Medical Concepts - GRIUM@CLEF2015 eHealth Task 2. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2015)
26. Oh, H.S., Jung, Y., Kim, K.Y.: KISTI at CLEF eHealth 2015 Task 2. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2015)
27. Thesprasith, O., Jaruskulchai, C.: Task 2a: Team KU-CS: Query Coherence Analysis for PRF and Genomics Expansion. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2015)
28. D'hondt, E., Grau, B., Zweigenbaum, P.: LIMSI @ CLEF eHealth 2015 - task 2. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2015)
29. Ksentini, N., Tmar, M., Boughanem, M., Gargouri, F.: Miracl at Clef 2015 : User-Centred Health Information Retrieval Task. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2015)
30. Huynh, N., Nguyen, T.T., Ho, Q.: TeamHCMUS: A Concept-Based Information Retrieval Approach for Web Medical Documents. In: Proceedings of the ShARe/-CLEF eHealth Evaluation Lab. (2015)
31. Thuma, E., Anderson, G., Mosweunyane, G.: UBML participation to CLEF eHealth IR challenge 2015: Task 2. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2015)
32. Lu, F.: Employing query expansion models to help patients diagnose themselves. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2015)
33. Ghoddousi, A., Huang, J.X., Feng, T.: York University at CLEF eHealth 2015: Medical Document Retrieval. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2015)

**Table 5.** Participants and baseline results sorted by p@10.

| R | Run Name | p@10 | nDCG@10 | R | Run Name | p@10 | nDCG@10 |
|---|---|---|---|---|---|---|---|
| **1** | **ECNU_EN_Run.3** | **0.5394** | **0.5086** | **55** | *readability_run.6* | *0.2970* | *0.2456* |
| **2** | ECNU_EN_Run.10 | 0.4667 | 0.4525 | **57** | Miracl_EN_Run.5 | 0.2939 | 0.2465 |
| **3** | ECNU_EN_Run.8 | 0.4530 | 0.4226 | **57** | YorkU_EN_Run.8 | 0.2939 | 0.2729 |
| **4** | ECNU_EN_Run.6 | 0.4227 | 0.3978 | **59** | YorkU_EN_Run.2 | 0.2924 | 0.2714 |
| **5** | **KISTI_EN_RUN.6** | **0.3864** | **0.3464** | **59** | YorkU_EN_Run.4 | 0.2924 | 0.2717 |
| **5** | KISTI_EN_RUN.8 | 0.3864 | 0.3464 | **59** | YorkU_EN_Run.6 | 0.2924 | 0.2694 |
| **7** | **CUNI_EN_Run.7** | **0.3803** | **0.3465** | **62** | FDUSGInfo_EN_Run.4 | 0.2848 | 0.2687 |
| **8** | KISTI_EN_RUN.4 | 0.3788 | 0.3424 | **62** | *baseline_run.4* | *0.2848* | *0.3483* |
| **9** | CUNI_EN_Run.4 | 0.3742 | 0.3409 | **64** | FDUSGInfo_EN_Run.5 | 0.2803 | 0.2665 |
| **10** | KISTI_EN_RUN.7 | 0.3727 | 0.3459 | **64** | YorkU_EN_Run.3 | 0.2803 | 0.2719 |
| **11** | CUNI_EN_Run.1 | 0.3712 | 0.3423 | **66** | YorkU_EN_Run.9 | 0.2788 | 0.2637 |
| **11** | CUNI_EN_Run.2 | 0.3712 | 0.3351 | **67** | UBML_EN_Run.5 | 0.2773 | 0.2500 |
| **13** | ECNU_EN_Run.4 | 0.3652 | 0.3168 | **68** | UBML_EN_Run.4 | 0.2742 | 0.2460 |
| **14** | **HCMUS_EN_Run.1** | **0.3636** | **0.3323** | **69** | USST_EN_Run.4 | 0.2727 | 0.2305 |
| **15** | CUNI_EN_Run.8 | 0.3621 | 0.3383 | **70** | UBML_EN_Run.9 | 0.2697 | 0.2538 |
| **16** | CUNI_EN_Run.6 | 0.3606 | 0.3364 | **70** | YorkU_EN_Run.10 | 0.2667 | 0.2546 |
| **16** | ECNU_EN_Run.2 | 0.3606 | 0.3220 | **72** | UBML_EN_Run.8 | 0.2652 | 0.2533 |
| **16** | ECNU_EN_Run.9 | 0.3606 | 0.3203 | **73** | **LIMSI_EN_run.3** | **0.2621** | **0.1960** |
| **16** | KISTI_EN_RUN.1 | 0.3606 | 0.3352 | **73** | UBML_EN_Run.6 | 0.2621 | 0.2265 |
| **16** | KISTI_EN_RUN.5 | 0.3606 | 0.3362 | **73** | *baseline_run.6* | *0.2621* | *0.3123* |
| **16** | ***readability_run.2*** | ***0.3606*** | ***0.3299*** | **76** | FDUSGInfo_EN_Run.2 | 0.2606 | 0.2488 |
| **22** | KISTI_EN_RUN.3 | 0.3591 | 0.3395 | **76** | HCMUS_EN_Run.3 | 0.2606 | 0.2341 |
| **23** | CUNI_EN_Run.5 | 0.3530 | 0.3217 | **78** | **KUCS_EN_Run.1** | **0.2545** | **0.2205** |
| **23** | CUNI_EN_Run.9 | 0.3530 | 0.3215 | **79** | Miracl_EN_Run.3 | 0.2515 | 0.1833 |
| **25** | CUNI_EN_Run.3 | 0.3485 | 0.3138 | **80** | UBML_EN_Run.10 | 0.2485 | 0.2294 |
| **26** | ECNU_EN_Run.1 | 0.3470 | 0.3144 | **81** | USST_EN_Run.5 | 0.2470 | 0.2082 |
| **27** | KISTI_EN_RUN.2 | 0.3455 | 0.3223 | **81** | USST_EN_Run.6 | 0.2470 | 0.2056 |
| **28** | *readability_run.1* | *0.3424* | *0.3226* | **83** | USST_EN_Run.7 | 0.2439 | 0.2220 |
| **29** | **USST_EN_Run.2** | **0.3379** | **0.3000** | **84** | Miracl_EN_Run.2 | 0.2424 | 0.1965 |
| **30** | *readability_run.3* | *0.3364* | *0.2890* | **85** | FDUSGInfo_EN_Run.3 | 0.2348 | 0.2234 |
| **31** | HCMUS_EN_Run.2 | 0.3348 | 0.3137 | **86** | LIMSI_EN_run.1 | 0.2318 | 0.1801 |
| **32** | ***baseline_run.1*** | ***0.3333*** | ***0.3151*** | **87** | LIMSI_EN_run.2 | 0.2303 | 0.1675 |
| **33** | *baseline_run.3* | *0.3242* | *0.2960* | **88** | KUCS_EN_Run.2 | 0.2288 | 0.1980 |
| **34** | ECNU_EN_Run.7 | 0.3227 | 0.3004 | **88** | *readability_run.7* | *0.2288* | *0.1834* |
| **35** | **Miracl_EN_Run.1** | **0.3212** | **0.2787** | **90** | USST_EN_Run.8 | 0.1985 | 0.1757 |
| **36** | **UBML_EN_Run.2** | **0.3197** | **0.2909** | **91** | HCMUS_EN_Run.4 | 0.1955 | 0.1866 |
| **37** | **GRIUM_EN_Run.6** | **0.3182** | **0.2944** | **91** | *baseline_run.5* | *0.1955* | *0.2417* |
| **37** | UBML_EN_Run.3 | 0.3182 | 0.2919 | **93** | Miracl_EN_Run.4 | 0.1894 | 0.1572 |
| **39** | GRIUM_EN_Run.3 | 0.3167 | 0.2913 | **93** | YorkU_EN_Run.1 | 0.1894 | 0.1718 |
| **40** | ECNU_EN_Run.5 | 0.3152 | 0.3006 | **95** | HCMUS_EN_Run.5 | 0.1545 | 0.1574 |
| **41** | GRIUM_EN_Run.1 | 0.3136 | 0.2875 | **96** | HCMUS_EN_Run.7 | 0.1470 | 0.1550 |
| **42** | UBML_EN_Run.1 | 0.3106 | 0.2897 | **97** | USST_EN_Run.9 | 0.1439 | 0.1241 |
| **43** | GRIUM_EN_Run.2 | 0.3091 | 0.2850 | **98** | USST_EN_Run.10 | 0.1348 | 0.1145 |
| **43** | UBML_EN_Run.7 | 0.3091 | 0.2887 | **99** | *readability_run.4* | *0.1227* | *0.0958* |
| **45** | *readability_run.5* | *0.3076* | *0.2595* | **100** | HCMUS_EN_Run.6 | 0.1045 | 0.1139 |
| **46** | GRIUM_EN_Run.7 | 0.3061 | 0.2798 | **101** | HCMUS_EN_Run.8 | 0.0970 | 0.1078 |
| **47** | GRIUM_EN_Run.5 | 0.3045 | 0.2803 | **102** | FDUSGInfo_EN_Run.6 | 0.0773 | 0.0708 |
| **47** | USST_EN_Run.1 | 0.3045 | 0.2841 | **102** | FDUSGInfo_EN_Run.7 | 0.0773 | 0.0708 |
| **49** | GRIUM_EN_Run.4 | 0.3030 | 0.2788 | **102** | FDUSGInfo_EN_Run.8 | 0.0773 | 0.0708 |
| **49** | USST_EN_Run.3 | 0.3030 | 0.2627 | **105** | FDUSGInfo_EN_Run.9 | 0.0682 | 0.0602 |
| **51** | **YorkU_EN_Run.7** | **0.3015** | **0.2766** | **105** | FDUSGInfo_EN_Run.10 | 0.0682 | 0.0602 |
| **51** | *baseline_run.2* | *0.3015* | *0.2479* | **107** | LIMSI_EN_run.4 | 0.0561 | 0.0378 |
| **53** | CUNI_EN_Run.10 | 0.3000 | 0.2597 | **107** | LIMSI_EN_run.5 | 0.0561 | 0.0378 |
| **53** | YorkU_EN_Run.5 | 0.3000 | 0.2752 | **109** | KUCS_EN_Run.3 | 0.0364 | 0.0299 |
| **55** | **FDUSGInfo_EN_Run.1** | **0.2970** | **0.2718** | **110** | KUCS_EN_Run.4 | 0.0182 | 0.0163 |

**Table 6.** Participants and baseline results sorted by uRBP.

| R | Run Name | RBP | uRBP | uRBPgr | R | Run Name | RBP | uRBP | uRBPgr |
|---|---|---|---|---|---|---|---|---|---|
| **1** | **ECNU_EN_Run.3** | **0.5339** | **0.3877** | **0.4046** | **56** | FDUSGInfo_EN_Run.4 | 0.3019 | 0.2373 | 0.2393 |
| **2** | ECNU_EN_Run.10 | 0.4955 | 0.3768 | 0.3873 | **57** | YorkU_EN_Run.6 | 0.3081 | 0.2365 | 0.2431 |
| **3** | **CUNI_EN_Run.7** | **0.3946** | **0.3422** | **0.3312** | **58** | YorkU_EN_Run.5 | 0.3109 | 0.2357 | 0.2416 |
| **4** | ECNU_EN_Run.6 | 0.4459 | 0.3374 | 0.3453 | **59** | UBML_EN_Run.8 | 0.2978 | 0.2352 | 0.2368 |
| **5** | CUNI_EN_Run.2 | 0.3796 | 0.3354 | 0.3239 | **60** | UBML_EN_Run.6 | 0.2766 | 0.2348 | 0.2310 |
| **6** | CUNI_EN_Run.5 | 0.3736 | 0.3295 | 0.3169 | **61** | FDUSGInfo_EN_Run.5 | 0.2989 | 0.2340 | 0.2356 |
| **7** | CUNI_EN_Run.9 | 0.3727 | 0.3287 | 0.3163 | **62** | YorkU_EN_Run.2 | 0.3151 | 0.2334 | 0.2404 |
| **8** | CUNI_EN_Run.4 | 0.3894 | 0.3284 | 0.3256 | **63** | UBML_EN_Run.9 | 0.2993 | 0.2332 | 0.2362 |
| **9** | ECNU_EN_Run.8 | 0.4472 | 0.3273 | 0.3373 | **64** | YorkU_EN_Run.4 | 0.3152 | 0.2319 | 0.2397 |
| **10** | ECNU_EN_Run.9 | 0.3730 | 0.3249 | 0.3107 | **65** | **KUCS_EN_Run.1** | **0.2785** | **0.2312** | **0.2251** |
| **11** | CUNI_EN_Run.6 | 0.3779 | 0.3224 | 0.3152 | **66** | *baseline_run.4* | *0.3196* | *0.2291* | *0.2323* |
| **12** | CUNI_EN_Run.3 | 0.3650 | 0.3218 | 0.3110 | **67** | Miracl_EN_Run.5 | 0.2982 | 0.2262 | 0.2357 |
| **13** | ***readability_run.2*** | ***0.3756*** | ***0.3154*** | ***0.3117*** | ***68*** | UBML_EN_Run.4 | 0.2953 | 0.2255 | 0.2300 |
| **14** | *readability_run.1* | *0.3675* | *0.3140* | *0.3064* | **69** | FDUSGInfo_EN_Run.2 | 0.2757 | 0.2237 | 0.2252 |
| **15** | ECNU_EN_Run.4 | 0.3638 | 0.3103 | 0.2990 | **70** | UBML_EN_Run.5 | 0.2960 | 0.2220 | 0.2279 |
| **16** | ECNU_EN_Run.1 | 0.3549 | 0.3080 | 0.2971 | **71** | YorkU_EN_Run.3 | 0.3074 | 0.2216 | 0.2300 |
| **17** | *readability_run.3* | *0.3390* | *0.3067* | *0.2929* | **72** | USST_EN_Run.3 | 0.3148 | 0.2181 | 0.2336 |
| **18** | CUNI_EN_Run.8 | 0.3842 | 0.3060 | 0.3102 | **73** | UBML_EN_Run.10 | 0.2658 | 0.2125 | 0.2159 |
| **19** | CUNI_EN_Run.1 | 0.3824 | 0.3027 | 0.3081 | **74** | FDUSGInfo_EN_Run.3 | 0.2518 | 0.2114 | 0.2087 |
| **20** | **HCMUS_EN_Run.1** | **0.3715** | **0.3017** | **0.3062** | **75** | USST_EN_Run.7 | 0.2726 | 0.2055 | 0.2102 |
| **21** | ***baseline_run.1*** | ***0.3567*** | ***0.2990*** | ***0.2933*** | **76** | **LIMSI_EN_run.3** | **0.2417** | **0.2036** | **0.2060** |
| **22** | ECNU_EN_Run.2 | 0.3527 | 0.2917 | 0.2830 | **77** | *baseline_run.6* | *0.2843* | *0.2035* | *0.2143* |
| **23** | ECNU_EN_Run.7 | 0.3548 | 0.2841 | 0.2869 | **78** | HCMUS_EN_Run.3 | 0.2700 | 0.2012 | 0.2089 |
| **24** | **GRIUM_EN_Run.2** | **0.3305** | **0.2809** | **0.2768** | **79** | USST_EN_Run.4 | 0.2815 | 0.1978 | 0.2110 |
| **25** | **UBML_EN_Run.7** | **0.3339** | **0.2795** | **0.2772** | **80** | LIMSI_EN_run.1 | 0.2296 | 0.1929 | 0.1889 |
| **26** | GRIUM_EN_Run.6 | 0.3306 | 0.2791 | 0.2761 | **81** | KUCS_EN_Run.2 | 0.2562 | 0.1818 | 0.1906 |
| **27** | GRIUM_EN_Run.5 | 0.3278 | 0.2780 | 0.2744 | **82** | LIMSI_EN_run.2 | 0.2163 | 0.1815 | 0.1774 |
| **28** | GRIUM_EN_Run.4 | 0.3244 | 0.2778 | 0.2719 | **83** | USST_EN_Run.5 | 0.2540 | 0.1746 | 0.1890 |
| **29** | GRIUM_EN_Run.3 | 0.3296 | 0.2775 | 0.2745 | **84** | Miracl_EN_Run.3 | 0.2200 | 0.1698 | 0.1698 |
| **30** | GRIUM_EN_Run.7 | 0.3272 | 0.2774 | 0.2739 | **85** | USST_EN_Run.6 | 0.2410 | 0.1633 | 0.1771 |
| **31** | ECNU_EN_Run.5 | 0.3531 | 0.2771 | 0.2804 | **86** | Miracl_EN_Run.2 | 0.2291 | 0.1589 | 0.1626 |
| **32** | UBML_EN_Run.3 | 0.3358 | 0.2757 | 0.2789 | **87** | *baseline_run.5* | *0.2226* | *0.1530* | *0.1610* |
| **33** | UBML_EN_Run.1 | 0.3294 | 0.2745 | 0.2771 | **88** | KUCS_EN_Run.3 | 0.1679 | 0.1514 | 0.1425 |
| **34** | *baseline_run.3* | *0.3369* | *0.2736* | *0.2751* | **89** | Miracl_EN_Run.4 | 0.2001 | 0.1507 | 0.1570 |
| **35** | GRIUM_EN_Run.1 | 0.3249 | 0.2725 | 0.2700 | **90** | USST_EN_Run.8 | 0.2246 | 0.1492 | 0.1595 |
| **36** | UBML_EN_Run.2 | 0.3305 | 0.2709 | 0.2735 | **91** | HCMUS_EN_Run.4 | 0.2099 | 0.1467 | 0.1582 |
| **37** | **KISTI_EN_RUN.7** | **0.3299** | **0.2703** | **0.2739** | **92** | HCMUS_EN_Run.5 | 0.1861 | 0.1299 | 0.1386 |
| **38** | KISTI_EN_RUN.5 | 0.3203 | 0.2702 | 0.2725 | **93** | HCMUS_EN_Run.7 | 0.1853 | 0.1266 | 0.1348 |
| **39** | **USST_EN_Run.2** | **0.3557** | **0.2659** | **0.2727** | **94** | YorkU_EN_Run.1 | 0.1798 | 0.1127 | 0.1195 |
| **40** | KISTI_EN_RUN.4 | 0.3306 | 0.2644 | 0.2709 | **95** | USST_EN_Run.9 | 0.1629 | 0.1115 | 0.1195 |
| **41** | *baseline_run.2* | *0.3150* | *0.2633* | *0.2587* | **96** | *readability_run.4* | *0.1143* | *0.1080* | *0.1000* |
| **42** | KISTI_EN_RUN.6 | 0.3332 | 0.2607 | 0.2695 | **97** | USST_EN_Run.10 | 0.1467 | 0.0947 | 0.1039 |
| **42** | KISTI_EN_RUN.8 | 0.3332 | 0.2607 | 0.2695 | **98** | HCMUS_EN_Run.6 | 0.1257 | 0.0746 | 0.0861 |
| **42** | KISTI_EN_RUN.2 | 0.3038 | 0.2607 | 0.2614 | **99** | HCMUS_EN_Run.8 | 0.1210 | 0.0698 | 0.0808 |
| **45** | KISTI_EN_RUN.3 | 0.3295 | 0.2596 | 0.2666 | **100** | FDUSGInfo_EN_Run.6 | 0.0805 | 0.0609 | 0.0577 |
| **46** | KISTI_EN_RUN.1 | 0.3222 | 0.2593 | 0.2646 | **100** | FDUSGInfo_EN_Run.7 | 0.0805 | 0.0609 | 0.0577 |
| **47** | **FDUSGInfo_EN_Run.1** | **0.3134** | **0.2572** | **0.2568** | **100** | FDUSGInfo_EN_Run.8 | 0.0805 | 0.0609 | 0.0577 |
| **48** | USST_EN_Run.1 | 0.3342 | 0.2564 | 0.2639 | **103** | KUCS_EN_Run.4 | 0.0656 | 0.0600 | 0.0567 |
| **49** | HCMUS_EN_Run.2 | 0.3483 | 0.2556 | 0.2698 | **104** | LIMSI_EN_run.4 | 0.0562 | 0.0476 | 0.0462 |
| **50** | **Miracl_EN_Run.1** | **0.3287** | **0.2546** | **0.2631** | **104** | LIMSI_EN_run.5 | 0.0562 | 0.0476 | 0.0462 |
| **51** | **YorkU_EN_Run.8** | **0.3072** | **0.2504** | **0.2533** | **106** | FDUSGInfo_EN_Run.9 | 0.0646 | 0.0473 | 0.0473 |
| **52** | YorkU_EN_Run.7 | 0.3125 | 0.2470 | 0.2523 | **107** | FDUSGInfo_EN_Run.10 | 0.0646 | 0.0473 | 0.0473 |
| **52** | YorkU_EN_Run.9 | 0.2962 | 0.2470 | 0.2485 | **108** | *readability_run.5* | *0.0362* | *0.0160* | *0.0227* |
| **54** | CUNI_EN_Run.10 | 0.3060 | 0.2442 | 0.2459 | **109** | *readability_run.6* | *0.0194* | *0.0117* | *0.0134* |
| **55** | YorkU_EN_Run.10 | 0.2853 | 0.2415 | 0.2420 | **109** | *readability_run.7* | *0.0194* | *0.0117* | *0.0134* |

**Table 7.** Results for multilingual submissions, sorted by p@10, obtained using the original qrels.

| R | Run Name | p@10 | nDCG@10 | R | Run Name | p@10 | nDCG@10 |
|---|---|---|---|---|---|---|---|
| **1** | CUNI_DE_Run10 | 0.2985 | 0.2825 | **34** | CUNI_IT_Run5 | 0.2182 | 0.1856 |
| **2** | CUNI_DE_Run7 | 0.2970 | 0.2757 | **37** | CUNI_AR_Run1 | 0.2167 | 0.2117 |
| **3** | CUNI_FR_Run10 | 0.2833 | 0.2615 | **37** | CUNI_AR_Run7 | 0.2167 | 0.2133 |
| **4** | CUNI_FR_Run7 | 0.2773 | 0.2568 | **39** | CUNI_CS_Run8 | 0.2152 | 0.2137 |
| **5** | CUNI_IT_Run10 | 0.2758 | 0.2369 | **39** | CUNI_FR_Run1 | 0.2152 | 0.2056 |
| **6** | CUNI_IT_Run1 | 0.2652 | 0.2278 | **39** | CUNI_PT_Run2 | 0.2152 | 0.2227 |
| **7** | CUNI_IT_Run4 | 0.2621 | 0.2221 | **42** | CUNI_FA_Run6 | 0.2136 | 0.2107 |
| **8** | CUNI_PT_Run6 | 0.2530 | 0.2492 | **43** | CUNI_CS_Run1 | 0.2121 | 0.1924 |
| **9** | CUNI_PT_Run8 | 0.2515 | 0.2382 | **43** | CUNI_DE_Run1 | 0.2121 | 0.1969 |
| **10** | CUNI_DE_Run8 | 0.2500 | 0.2413 | **43** | CUNI_IT_Run7 | 0.2121 | 0.1812 |
| **10** | CUNI_FR_Run9 | 0.2500 | 0.2188 | **43** | CUNI_PT_Run3 | 0.2121 | 0.2253 |
| **12** | CUNI_FR_Run8 | 0.2455 | 0.2271 | **47** | CUNI_FA_Run7 | 0.2091 | 0.1806 |
| **12** | CUNI_IT_Run6 | 0.2455 | 0.2142 | **48** | CUNI_CS_Run5 | 0.2076 | 0.1958 |
| **14** | CUNI_DE_Run9 | 0.2409 | 0.2107 | **48** | CUNI_FR_Run3 | 0.2076 | 0.1943 |
| **14** | CUNI_PT_Run10 | 0.2409 | 0.2451 | **48** | CUNI_FR_Run5 | 0.2076 | 0.2017 |
| **16** | CUNI_IT_Run2 | 0.2394 | 0.1913 | **51** | CUNI_FR_Run4 | 0.2061 | 0.2074 |
| **17** | CUNI_IT_Run3 | 0.2348 | 0.1952 | **52** | CUNI_AR_Run8 | 0.2045 | 0.2026 |
| **17** | CUNI_PT_Run7 | 0.2348 | 0.2266 | **52** | CUNI_DE_Run5 | 0.2045 | 0.1940 |
| **19** | CUNI_IT_Run8 | 0.2333 | 0.2105 | **54** | CUNI_AR_Run4 | 0.2030 | 0.1966 |
| **20** | CUNI_CS_Run10 | 0.2303 | 0.1926 | **54** | CUNI_AR_Run9 | 0.2030 | 0.1768 |
| **20** | CUNI_FA_Run10 | 0.2303 | 0.2277 | **54** | CUNI_CS_Run6 | 0.2030 | 0.1605 |
| **20** | CUNI_PT_Run1 | 0.2303 | 0.2338 | **57** | CUNI_DE_Run4 | 0.2015 | 0.1869 |
| **20** | CUNI_PT_Run5 | 0.2303 | 0.2180 | **58** | CUNI_DE_Run3 | 0.2000 | 0.1652 |
| **24** | CUNI_PT_Run4 | 0.2288 | 0.2352 | **59** | CUNI_FA_Run9 | 0.1985 | 0.1735 |
| **25** | CUNI_AR_Run10 | 0.2273 | 0.2202 | **60** | CUNI_FR_Run6 | 0.1970 | 0.1661 |
| **25** | CUNI_FA_Run4 | 0.2273 | 0.2267 | **61** | CUNI_CS_Run9 | 0.1924 | 0.1530 |
| **25** | CUNI_IT_Run9 | 0.2273 | 0.1856 | **62** | CUNI_CS_Run4 | 0.1894 | 0.1721 |
| **28** | CUNI_FA_Run1 | 0.2258 | 0.2227 | **63** | CUNI_AR_Run2 | 0.1879 | 0.1831 |
| **29** | CUNI_CS_Run7 | 0.2242 | 0.1897 | **63** | CUNI_FR_Run2 | 0.1879 | 0.1854 |
| **30** | CUNI_FA_Run3 | 0.2227 | 0.2049 | **63** | CUNI_PT_Run9 | 0.1879 | 0.1719 |
| **30** | CUNI_FA_Run5 | 0.2227 | 0.1991 | **66** | CUNI_AR_Run3 | 0.1864 | 0.1894 |
| **32** | CUNI_AR_Run5 | 0.2197 | 0.2148 | **67** | CUNI_CS_Run3 | 0.1848 | 0.1609 |
| **32** | CUNI_AR_Run6 | 0.2197 | 0.2017 | **68** | CUNI_DE_Run6 | 0.1818 | 0.1485 |
| **34** | CUNI_FA_Run2 | 0.2182 | 0.2087 | **69** | CUNI_CS_Run2 | 0.1697 | 0.1470 |
| **34** | CUNI_FA_Run8 | 0.2182 | 0.2201 | **69** | CUNI_DE_Run2 | 0.1697 | 0.1517 |

**Table 8.** Results for multilingual submissions, sorted by p@10, obtained using additional qrels (*merged*).

| R | Run Name | p@10 | nDCG@10 | R | Run Name | p@10 | nDCG@10 |
|---|---|---|---|---|---|---|---|
| **1** | CUNI_IT_Run10 | 0.3727 | 0.3094 | **36** | CUNI_FR_Run2 | 0.3061 | 0.2498 |
| **1** | CUNI_IT_Run4 | 0.3727 | 0.3045 | **37** | CUNI_FA_Run5 | 0.3045 | 0.2539 |
| **3** | CUNI_IT_Run1 | 0.3712 | 0.3065 | **38** | CUNI_AR_Run5 | 0.3030 | 0.2661 |
| **4** | CUNI_FR_Run10 | 0.3682 | 0.3111 | **38** | CUNI_CS_Run3 | 0.3030 | 0.2259 |
| **4** | CUNI_FR_Run7 | 0.3682 | 0.3093 | **38** | CUNI_CS_Run4 | 0.3030 | 0.2343 |
| **6** | CUNI_IT_Run6 | 0.3606 | 0.2981 | **41** | CUNI_CS_Run6 | 0.3000 | 0.2206 |
| **7** | CUNI_PT_Run2 | 0.3576 | 0.3009 | **41** | CUNI_FA_Run8 | 0.3000 | 0.2669 |
| **8** | CUNI_DE_Run10 | 0.3561 | 0.3182 | **43** | CUNI_DE_Run8 | 0.2985 | 0.2672 |
| **9** | CUNI_DE_Run7 | 0.3545 | 0.3092 | **43** | CUNI_PT_Run7 | 0.2985 | 0.2613 |
| **10** | CUNI_IT_Run8 | 0.3515 | 0.2966 | **45** | CUNI_AR_Run10 | 0.2924 | 0.2556 |
| **11** | CUNI_PT_Run1 | 0.3500 | 0.2936 | **45** | CUNI_AR_Run7 | 0.2924 | 0.2569 |
| **12** | CUNI_PT_Run4 | 0.3485 | 0.2976 | **45** | CUNI_CS_Run8 | 0.2924 | 0.2544 |
| **13** | CUNI_IT_Run2 | 0.3424 | 0.2683 | **45** | CUNI_DE_Run9 | 0.2924 | 0.2397 |
| **14** | CUNI_IT_Run3 | 0.3394 | 0.2694 | **45** | CUNI_PT_Run9 | 0.2924 | 0.2255 |
| **15** | CUNI_PT_Run10 | 0.3379 | 0.2936 | **50** | CUNI_CS_Run5 | 0.2909 | 0.2426 |
| **16** | CUNI_PT_Run3 | 0.3364 | 0.2893 | **51** | CUNI_AR_Run6 | 0.2894 | 0.2392 |
| **17** | CUNI_FA_Run10 | 0.3333 | 0.2807 | **52** | CUNI_AR_Run8 | 0.2879 | 0.2493 |
| **17** | CUNI_FR_Run9 | 0.3333 | 0.2677 | **52** | CUNI_FR_Run5 | 0.2879 | 0.2446 |
| **17** | CUNI_PT_Run6 | 0.3333 | 0.2893 | **54** | CUNI_CS_Run9 | 0.2864 | 0.2122 |
| **20** | CUNI_CS_Run1 | 0.3318 | 0.2633 | **54** | CUNI_FA_Run6 | 0.2864 | 0.2504 |
| **20** | CUNI_CS_Run7 | 0.3318 | 0.2571 | **56** | CUNI_FA_Run7 | 0.2803 | 0.2256 |
| **20** | CUNI_IT_Run5 | 0.3318 | 0.2666 | **56** | CUNI_FR_Run6 | 0.2803 | 0.2070 |
| **20** | CUNI_IT_Run7 | 0.3318 | 0.2654 | **58** | CUNI_DE_Run1 | 0.2773 | 0.2327 |
| **20** | CUNI_PT_Run8 | 0.3318 | 0.2838 | **59** | CUNI_DE_Run4 | 0.2742 | 0.2255 |
| **25** | CUNI_CS_Run10 | 0.3288 | 0.2567 | **60** | CUNI_AR_Run1 | 0.2727 | 0.2403 |
| **26** | CUNI_FA_Run4 | 0.3273 | 0.2788 | **60** | CUNI_CS_Run2 | 0.2727 | 0.2058 |
| **26** | CUNI_FR_Run3 | 0.3273 | 0.2612 | **60** | CUNI_FA_Run9 | 0.2727 | 0.2101 |
| **28** | CUNI_FA_Run1 | 0.3258 | 0.2720 | **63** | CUNI_DE_Run3 | 0.2682 | 0.2039 |
| **29** | CUNI_FA_Run3 | 0.3242 | 0.2660 | **64** | CUNI_AR_Run2 | 0.2621 | 0.2178 |
| **30** | CUNI_FA_Run2 | 0.3227 | 0.2674 | **64** | CUNI_AR_Run4 | 0.2621 | 0.2237 |
| **31** | CUNI_FR_Run4 | 0.3182 | 0.2661 | **64** | CUNI_DE_Run5 | 0.2621 | 0.2211 |
| **31** | CUNI_FR_Run8 | 0.3182 | 0.2659 | **67** | CUNI_AR_Run3 | 0.2591 | 0.2261 |
| **31** | CUNI_PT_Run5 | 0.3182 | 0.2717 | **68** | CUNI_DE_Run2 | 0.2485 | 0.1984 |
| **34** | CUNI_FR_Run1 | 0.3121 | 0.2557 | **69** | CUNI_AR_Run9 | 0.2439 | 0.1954 |
| **35** | CUNI_IT_Run9 | 0.3106 | 0.2417 | **70** | CUNI_DE_Run6 | 0.2364 | 0.1811 |