The Covid-19 MLIA @ Eval Initiative: Overview of the Multilingual Semantic Search Task

Giorgio Maria Di Nunzio[1], Maria Eskevich[2], and Nicola Ferro[1]

[1] University of Padua, Italy
giorgiomaria.dinunzio@unipd.it
nicola.ferro@unipd.it
[2] CLARIN ERIC, The Netherlands
maria@clarin.eu

Abstract. This report provides an overview of the first round of Task 2. For each language, we give an account of the size of the corpora used for the two subtasks, and we describe the structure of the thirty topics as well as the creation of the pool for the relevance judgements. We received from the four participants for subtask 1 a total of 60 mononlingual runs and 35 bilingual runs, while for subtask 2 a total of 49 monolingual runs and 31 bilingual runs.

1 Task description

The goal of the Multilingual Semantic Search task is to collect relevant information for the community, the general public as well as other stakeholders, when searching for health content in different languages and with different levels of knowledge about the specific topic. There are two sub-tasks:

- **Subtask 1** is a classic ad-hoc multilingual search task focused more on high precision.
- **Subtask 2** is more oriented towards high-recall systems, like Technology Assisted Review (TAR) systems.

For the first subtask, participants can submit a run with at most 1,000 documents per topic (for a total of 30,000 retrieved documents). For subtask 2, each run must have at most a total of 6,000 documents retrieved overall. In this second subtask, we expect on average 200 document retrieved per topic; however, participants can decide to distribute documents unevenly.

In the first round, the systems work without relevant information. From the second round, the systems can use the information about the relevance assessments to optimize their systems.

Covid-19 MLIA @ Eval Initiative, http://eval.covid19-mlia.eu/.

2 Collection

In this first round, eight languages, seven of which are official EU languages, were chosen to build the initial set of collections of documents. These languages are related to countries where COVID spread quickly or was managed in a different way at the beginning of 2020. For each language, five corpora where selected.

The details of corpora can be found in Table 2.

Language	Number of documents per corpora					TOTAL
	EU Press Corner	EUR-Lex	Global Voices	MEDISYS	Wikipedia	
English (en)	335	352	571	1 450 251	731	**1 452 240**
French (fr)	276	345	446	325 178	357	**326 599**
German (de)	266	345	51	272 645	364	**273 761**
Greek (el)	120	344	328	146 763	103	**147 658**
Italian (it)	123	342	539	661 514	271	**662 789**
Spanish (es)	115	342	595	832 639	342	**833 763**
Swedish (sv)	122	343	5	37 615	111	**38 196**
Ukrainian (uk)	0	0	66	15 395	121	**15 582**

Table 1. Overview of the corpora used for Round 1.

3 Set of topics

The topics were created by selecting 1) a subset of the queries created for the TREC-COVID Task[3] (courtesy of TREC-COVID Task organizers) [1] and 2) a selection of queries made available in the Bing search dataset for Coronavirus Intent[4] which includes queries from all over the world that had an explicit/implicit intent related to the Coronavirus or Covid-19.

Topics are structured in the following way:

```
<topic number"topic identifier" xml:lang="ISO 639-1 code" >
<keyword>keyword based query</keyword>
<conversational>the query as a question posed by the user
</conversational>
<explanation>a more detailed explanation of
what the set of retrieved documents should look like</explanation>
</topic>
```

The keyword field represents the "traditional" way a user performs the search on a Web search engine. It is basically a set of keywords, i.e. "surgical mask protection". The conversational field is more like a way of asking the same thing in a

[3] `https://ir.nist.gov/covidSubmit/`

[4] `https://github.com/microsoft/BingCoronavirusQuerySet`

Language	cunimtir	gatenlp	ims	sinai	total
English	5 / 20	5 / 0	5 / 0	0 / 0	15 / 20
French	0 / 0	5 / 5	4 / 0	0 / 0	9 / 5
German	0 / 0	5 / 5	4 / 0	0 / 0	9 / 5
Greek	0 / 0	0 / 0	3 / 0	0 / 0	3 / 0
Italian	0 / 0	0 / 0	4 / 0	0 / 0	4 / 0
Spanish	0 / 0	5 / 5	4 / 0	5 / 0	14 / 0
Swedish	0 / 0	0 / 0	3 / 0	0 / 0	3 / 0
Ukrainian	0 / 0	0 / 0	3 / 0	0 / 0	3 / 0
total	5 / 20	20 / 15	30 / 0	5 / 0	60 / 35

Table 2. Subtask 1. Submitted runs (monolingual/bilingual) per language and participants.

Language	cunimtir	gatenlp	ims	sinai	total
English	4 / 16	5 / 0	4 / 0	0 / 0	13 / 16
French	0 / 0	5 / 5	4 / 0	0 / 0	9 / 5
German	0 / 0	5 / 5	4 / 0	0 / 0	9 / 5
Greek	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
Italian	0 / 0	0 / 0	4 / 0	0 / 0	4 / 0
Spanish	0 / 0	5 / 5	4 / 0	5 / 0	14 / 5
Swedish	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
Ukrainian	0 / 0	0 / 0	0 / 0	0 / 0	0 / 0
total	4 / 16	20 / 15	20 / 0	5 / 0	49 / 31

Table 3. Subtask 2. Submitted runs (monolingual/bilingual) per language and participants.

verbal way, i.e. "does a surgical mask protect from Covid-19?" The explanation field is used to provide information to the assessors when performing relevance assessments, i.e. "The documents retrieved should contain information about ...".

These topics have been manually translated from English into 8 languages of the available corpora, as well as into Chinese (zh) and Japanese (ja).

4 Participants

Four participants submitted runs for this task (listed in order of user id):

- Charles University, Czech Republic (cunimtir);
- University of Sheffield, UK (gatenlp);
- University of Padua, Italy (ims);
- Universidad de Jaén, Spain (sinai).

In Table 2 and Table 3, we report the number of monolingual and bilingual runs submitted by each participant.

Language	mono sub 1	mono sub 2	bili sub 1	bili sub 2	total	achieved	# of assessors
English	15	15	5	5	8,247	7,242	8
French	45	45	10	10	7,169	4,360	2
German	45	45	10	10	6,913	5,183	1
Greek	100	-	-	-	4,908	4,324	10
Italian	75	75	-	-	8,720	7,680	7
Spanish	23	23	5	5	7,091	7,091	4
Swedish	100	-	-	-	5,445	5,445	4

Table 4. For each language, the threshold used to select the top k document of each run (per subtask and type). The total number of document and the actual number of documents judged. The last column shows the number of assessors available per language.

5 Ground truth creation

5.1 Pooling across submitted runs

In order to build the pool of documents to be judged, we selected a number of top k documents for each run in order to reach, for each language, a pool of a size around 6,000 - 8,000 documents.

In Table 4, we report the threshold k for each language, subtask and type (monolingual or bilingual) and the number of documents in the pool.

Given the short time constraints of Round 1, for some languages we had to reduce at some point in time the number of documents to assess in order to complete the judgements. In those cases (English, French, German, Greek), we decided to distribute the remaining documents in order to have at least the top 5 documents for each run (independently from the subtask or type) judged.

The pool for Ukrainian could not be finished by the end of round 1, and we have postponed the release of the pool later in January 2021.

5.2 Relevance judgement

In Table 5, we show the number of judged relevant and the number of documents judged for each topic and language.

All the topics have at least one relevant documents. In some cases, the distribution of documents to judge (and relevant documents) resulted uneven across a language since we had to reduce the pool on the fly (see for example some French and German topics with less than 100 documents to judge).

On average for Swedish and Italian there is lower percentage of relevant documents, 32% and 34 % respectively, while for others it varies between 45-59%. The extreme case being represented by three topics that have only 1 and 2 relevant documents for Swedish language, while for other languages this number vary between 10 and 76 topics with comparable number of overall retrieved and assessed documents:

- 1129: ‘göra eget handdesinfektionsmedel’/‘how to make hand sanitizer’

– 1135: ‘covid nedstängningsprotester’ / ‘covid lockdown protests’
– 1115: ‘amorteringsstöd och coronavirus’/‘mortgage assistance coronavirus’

	English		French		German		Greek		Italian		Spanish		Swedish	
topic	# rel	# docs	# rel	# docs	# rel	# docs	# rel	# docs	# rel	# docs	# rel	# docs	# rel	# docs
1	106	231	122	313	170	219	71	122	102	271	168	247	103	174
3	123	231	75	138	153	227	118	178	137	228	138	193	83	149
4	72	273	35	183	131	242	68	182	81	373	67	266	54	180
6	160	255	90	175	265	277	98	106	188	381	211	293	28	171
7	118	155	43	120	102	150	111	200	118	332	146	205	61	194
10	149	290	66	151	95	199	111	203	61	301	111	233	45	148
11	125	288	82	144	70	226	62	133	66	189	82	241	27	256
12	110	249	187	246	21	225	124	212	80	371	113	292	114	216
13	107	231	58	201	106	175	139	164	106	289	189	243	126	223
14	104	279	85	111	105	164	80	257	78	143	91	160	93	103
18	154	249	87	162	131	150	94	146	73	288	119	216	41	143
19	145	227	65	118	151	168	70	132	82	236	114	258	51	221
21	119	186	181	228	146	178	147	165	154	263	244	287	190	220
22	70	245	45	159	44	114	52	116	47	204	118	185	58	137
23	119	218	53	134	29	43	107	134	89	227	188	215	90	112
24	98	195	37	95	17	24	104	145	84	301	190	203	91	136
1101	111	236	94	224	157	280	70	107	74	316	160	253	182	251
1104	135	281	70	182	65	137	96	137	114	230	161	301	43	252
1105	77	251	62	232	63	243	27	114	25	181	53	219	31	123
1106	99	307	69	217	28	62	84	162	37	236	171	304	54	180
1110	68	316	112	182	51	146	59	136	85	275	157	271	14	153
1113	60	199	39	102	61	186	39	83	49	177	83	218	8	125
1115	173	294	54	120	32	38	124	152	65	207	213	247	2	259
1116	81	182	36	111	162	239	57	107	77	320	132	247	88	157
1120	173	223	32	40	100	123	72	109	111	262	182	246	29	231
1122	68	227	29	48	178	227	94	151	114	183	138	223	27	114
1123	95	148	12	19	93	150	21	100	38	191	126	155	9	176
1129	55	221	16	36	40	164	21	98	81	207	99	211	1	170
1130	144	216	110	124	101	122	39	119	135	253	140	193	49	205
1135	58	339	10	45	43	285	76	154	122	245	58	266	1	266

Table 5. Number of documents per language: relevance vs total.

6 Evaluation metrics

Subtask 1 has the main focus on the top ranked documents. Thus, the evaluation measures like Precision at 5 as well as Normalized Discounted Cumulative Gain are used to compare systems.
Subtask 2 focuses more on the problem of finding as many relevant documents as possible with the least effort. Given a limited amount of resources, such as a time limit and expert availability in time of crisis, there will be a limit on the maximum number of documents that can be retrieved in order to build a set of relevant documents that should be delivered to the general public. Evaluation measures like Precision@k and RPrec will be used to compare the systems.

7 Overview of the first round results

To be updated afer the Virtual Meeting of 12-14 January 2021.

References

[1] Voorhees, E.M., Alam, T., Bedrick, S., Demner-Fushman, D., Hersh, W.R., Lo, K., Roberts, K., Soboroff, I., Wang, L.L.: TREC-COVID: constructing a pandemic information retrieval test collection. CoRR **abs/2005.04474** (2020), URL `https://arxiv.org/abs/2005.04474`