# UNIPD at Covid-19 MLIA

Giorgio Maria Di Nunzio, Dennis Dosso, Alessandro Fabris, Guglielmo Faggioli, Nicola Ferro, Fabio Giachelle, Ornella Irrera, Stefano Marchesin, Luca Piazzon, Alberto Purpura,✉ Gianmaria Silvello, Federica Vezzani

University of Padua, Italy
purpuraa@dei.unipd.it

**Abstract.** In our participation to the first round of the Covid-19 MLIA Multilingual Semantic Search task, the UNIPD team focused on rapid development of information access tools, both from a system perspective and an evaluation perspective. On the system development side, we tested and compared different Information Retrieval (IR) approaches with little or no tuning. On the evaluation side, we set up a crowdsourcing experiment to assess the viability of quickly gathering relevance judgements from non-expert assessors in this information access scenario aimed at the general public.

## 1 Introduction

For the Covid-19 MLIA Multilingual Semantic Search task, the UNIPD team tested and compared different IR approaches. For the first round, in the absence of relevance judgements at development time, we considered off-the-shelf algorithms that required little tuning. Our rationale was to employ public resources, available to a practitioner in the early development stages of an information access tool.

In parallel, we set up a crowdsourcing experiment, emulating the process of quickly gathering relevance assessments to support supervised models. While we did not use crowdsourced judgements for model development during the first round, we carried out a comparison against official relevance assessments after they became available. Through this comparison we evaluated the possibility of quickly obtaining relevance judgements in a use case aimed at the general public.

Our work hinged on a flexible information processing pipeline, described in Section 2, combining diverse approaches to document preprocessing (Section 2.1) query reformulation (Section 2.2), document retrieval and ranking (Sections 2.3, 2.4), along with a final rank fusion stage (Section 2.5). The runs submitted for each language in the first round are detailed in Section 2.6 and evaluated in Section 3, where we also compare official and crowdsourced relevance judgements.

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

## 2 Proposed Approach

Our approach, summarized by the pipeline depicted in Figure 1, consisted of the following steps. We first processed the web pages of the given multi-lingual corpus to separate the document *body*, used as an input to our retrieval systems, from the remaining HTML tags (Section 2.1). In parallel, we performed query reformulation for English, French, German, Italian and Spanish over each topic in the collection (Section 2.2). Next, we computed different ranked lists using the query reformulations (where available) and several combinations of stoplists, stemmers and retrieval models. We employed both lexical models (Section 2.3) and neural models (Section 2.4). Finally, we used a rank fusion approach to merge the different runs (Section 2.5).

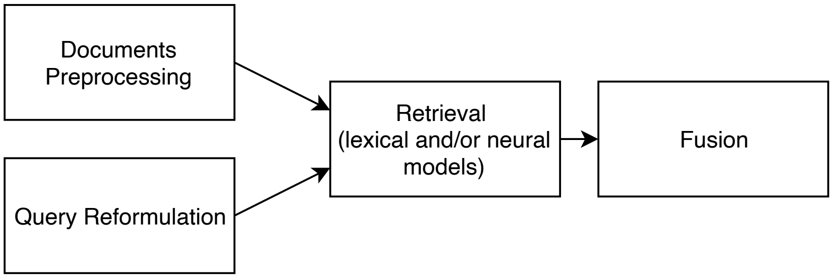


**Fig. 1.** Pipeline of the proposed approach.

### 2.1 Preprocessing

As a preprocessing step, we restricted the documents' content to the *body* field only and we removed 'boilerplate' information, as it can be a source of noise for the retrieval process. Finally, we stripped multiple white-spaces and we lowercased the *body* content.

### 2.2 Query Reformulation

In order to generate query variants from the original topics, we asked the students of the course Computer Assisted Translation Tools of the Master Degree in Modern Languages for International Communication and Cooperation of the University of Padua to focus on the medical terminology used in the Keyword version of the provided queries. First, they proceeded with the identification of single terms (ex. 'coronavirus') and multi-word terms (ex. 'contact tracing'). Once collected, the terms were analyzed in order to find terminological variations to represent the same medical concept. Different variants (semantic and/or orthographic) have been identified for each medical term. These equivalents were

| topic | query | variant |
|---|---|---|
| 1 | coronavirus origin | 0 |
| 1 | 2019 ncov origin | 1 |
| 1 | corona origin | 2 |
| 1 | covid origin | 3 |
| 1 | covid 19 origin | 4 |
| 1 | sars cov origin | 5 |

**Table 1.** Examples of query reformulations for the English topic 1. The variant "0" is the original topic.

| Language | # of reformulations |
|---|---|
| English | 1,139 |
| French | 225 |
| German | 414 |
| Italian | 1,632 |
| Spanish | 554 |

**Table 2.** Number of query reformulation per language.

used in order to replace the initially provided term and to generate query variants for each topic. In Table 1, we show an example of query variation for the query "coronavirus origin" (topic 1). A total of 82 students participated in this task for five languages (English, French, German, Italian, and Spanish), and we collected a few hundreds of query reformulations per language, as shown in Table2.

### 2.3 Lexical Models

Next, we built a grid of runs using traditional lexical approaches, summarized in Table 3. The search engine, based on the *Lucene* framework, combines different components in a Grid of Points (GoP) approach [2]. In a GoP, each component of the retrieval pipeline is used in combination with any possible combination of the remaining ones. More in detail, we considered three main components: the stoplist, the stemmer and the retrieval function. Our aim was to be as consistent as possible across languages.

Whenever possible, we used stemmers based on similar algorithms and stoplists built according to the same approach. We built the runs considering 4 publicly available stoplists, along with the no-stop approach, which does not apply any stoplist. More in detail, we employed the default Lucene stoplist, plus 3 stoplists available at `https://github.com/stopwords-iso/stopwords-iso`, for each supported language. As for stemmers, we considered two popular approaches, where available, plus the no-stem approach, which does not stem the tokens.

For each language, we used the same ranking functions, based on lexical Bag of Words (BoW) approaches. We considered the following retrieval functions: the okapi bm25 approach (bm25), language models with Dirichlet smoothing (lmd) or Jelinek-Mercer smoothing (lmjm), the tf-idf based approach, referred to as "classical" in Lucene (tf-idf), and Divergence From Randomness with Inverse Expected Document Frequency model with Bernoulli after-effect and normalization 2 (dfrinexpb2). Table 3 reports the main components used to build our lexical indexes, along with the final GoP size for each language.

For each point in the GoP and each topic, we computed three runs: one using only the title formulation of the topic, one combining the title and the conversational formulation of the topic, and one using the reformulations of queries (Section 2.2). For the first round of the initiative, we submitted two runs for each language, using the Lucene default stoplist (exception: ukStandard for Ukrainian), the lightest stemmer available (exceptions: greek stemmer for the Greek language, porter stemmer for English and nostem for Ukrainian), and the bm25 retrieval model.

**Table 3.** Stoplists, stemmers and ranking functions employed for each language. Stoplists marked with * are taken from `stopwords-iso`. Stoplists marked with † are the default stoplists in other search engines. The remaining components are available in Lucene after minor or no adaptation.

| lang. | stoplists | stemmers | ranking fun. | GoP size |
|---|---|---|---|---|
| **de** | bbalet*, ranksnl*, gh*, lucene, nostop | nostem, german, germanLight | bm25<br>tf-idf<br>lmd<br>lmjm<br>dfrinexpb2 | 75 |
| **el** | nostop, bbalet*, ranksnl*, gh*, lucene | nostem, greek | | 50 |
| **en** | nostop, lucene, indri†, atire†, okapi† | nostem, porter, lovins | | 75 |
| **es** | nostop, bbalet*, ranksnl*, gh*, lucene | nostem, spanishLight, snowball | | 75 |
| **fr** | nostop, bbalet*, ranksnl*, gh*, lucene | nostem, frenchLight, snowball | | 75 |
| **it** | nostop, bbalet*, ranksnl*, gh*, lucene | nostem, italianLight, snowball | | 75 |
| **sv** | nostop, fergiemcdowall*, bbalet*, gh*, lucene | nostem, swedishLight, snowball | | 75 |
| **uk** | nostop, ukrainianHeavy*, ranksnl*, ukStandard* | nostem | | 20 |

## 2.4 Neural Models

In parallel to lexical systems, we also considered neural models. We adopted SLEDGE [4], a search system that relies on SciBERT [1] to effectively re-rank scientific articles related to COVID-19. SLEDGE adopts a two-stage re-ranking pipeline to retrieve and rank documents. In the first stage, SLEDGE employs a traditional lexical model – such as bm25 [6] or Query Likelihood Model (QLM) [9] – to retrieve a recall-oriented set of candidate documents. In the second stage,

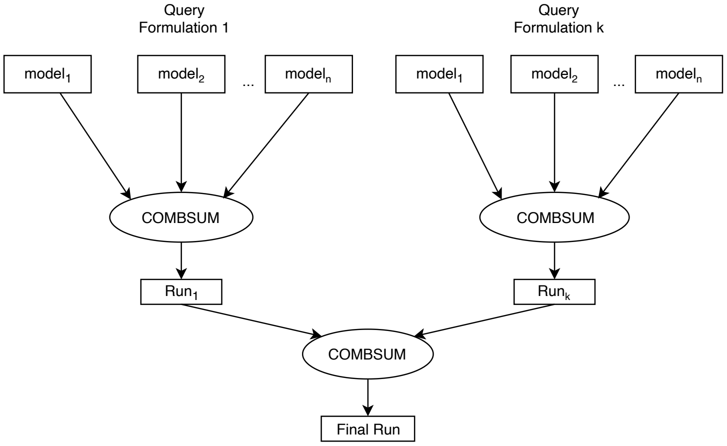


**Fig. 2.** Architecture of our rank fusion approach.

SLEDGE re-ranks the candidate documents through a SciBERT-based, high-precision neural ranking model. SLEDGE has been trained on the MS-MARCO dataset [5] – a general domain dataset that consists of over 800,000 (query, document) pairs and a shallow labeling scheme – and then transferred to the TREC-COVID dataset [8] for evaluation.

To perform retrieval on the Covid-19 MLIA dataset, we resorted to the two trained SLEDGE models available at `https://github.com/Georgetown-IR-Lab/covid-neural-ir`. The first model (sledge) has been trained on the whole MS-MARCO dataset, whereas the second (sledge-med) has been trained on the medical subset only. We relied on the OpenNIR library [3] to deploy the models on the Covid-19 MLIA dataset.[1] We set the bm25 parameters to $b = 0.75$ and $k1 = 1.2$, kept default parameters for SLEDGE, re-ranked the top 500 documents retrieved by bm25, and used the query reformulations described in Subsection 2.2.

### 2.5 Rank Fusion

The approach that we adopted to merge the different ranked lists consists of two stages, as schematized in Figure 2. The first stage, common to all languages, combines the runs from the lexical models (Section 2.3). In this stage, we consider all the retrieval pipeline combinations except the ones using the no-stem, tf-idf and lmjm variations. Indeed, according to our crowdsourcing experiment, the pipelines using these components are the worst performing overall. For languages where query reformulations (Section 2.2) are available, this fusion pro-

[1] `https://github.com/Georgetown-IR-Lab/OpenNIR/`

cess acts on each formulation independently. We performed rank fusion using the COMBSUM [7] algorithm[2] with a depth of 1000 documents.

The second stage is geared towards the languages for which query reformulations were available, namely English, French, German, Italian and Spanish. To merge runs associated with separate query reformulations, another fusion process is performed downstream of the first fusion step. We refer to the final run outputted by this two-stage fusion as `v-csum`. For the remaining languages (Ukrainian, Greek, Swedish), runs only underwent the first rank fusion stage, using the same configuration of the COMBSUM algorithm. The run obtained through this one-stage fusion is dubbed `csum`.

For the English language, we also exploited neural IR models. The outputs of sledge and sledge-med (Section 2.4) were fused in the first stage with a depth of 500 documents, while the second stage took care of combining the rankings obtained through different query reformulations. A third fusion stage combined the final neural run and the final lexical run into a single run dubbed `nlex`.

### 2.6 Submitted Runs

For each language, we submitted the following runs:

- `bm25`: bm25 with default Lucene $b$ and $k$ parameters, default Lucene stoplist. Queries coincide with the keyword-only formulation of each topic.
- `c-bm25`: same as above, with queries combining both keyword and conversational formulations.
- `csum`: one-stage fusion of all the lexical runs, using only the keyword-only formulation of the query.

For languages where query reformulations were available (English, French, German, Italian and Spanish), we also submitted the following run:

- `v-csum`: two-stage fusion, using all the available topic formulations and lexical runs, as described in Section 2.5.

For English, we did not submit the one-stage fusion run `csum`, in favour of:

- `nlex`: three-stage fusion, using all the available topic formulations, lexical runs and neural runs (Section 2.5).
- `nsle`: the output of sledge-med.

Table 4 contains the complete list of submitted runs along with details specific to each language.

[2] We relied on the COMBSUM implementation available at: `https://github.com/rmit-ir/polyfuse`. The parameters that we did not mention in the text were set to their default values.

**Table 4.** Submitted runs

| language | run | notes |
|---|---|---|
| **de** | `bm25` | germanLight stemmer |
| | `c-bm25` | germanLight stemmer |
| | `csum` | |
| | `v-csum` | |
| **el** | `bm25` | greek stemmer |
| | `c-bm25` | greek stemmer |
| | `csum` | |
| **en** | `bm25` | porter stemmer |
| | `c-bm25` | Porter stemmer |
| | `v-csum` | |
| | `nlex` | |
| | `nsle` | |
| **es** | `bm25` | spanishLight stemmer |
| | `c-bm25` | spanishLight stemmer |
| | `csum` | |
| | `v-csum` | |
| **fr** | `bm25` | frenchLight stemmer |
| | `c-bm25` | frenchLight stemmer |
| | `csum` | |
| | `v-csum` | |
| **it** | `bm25` | italianLight stemmer |
| | `c-bm25` | italianLight stemmer |
| | `csum` | |
| | `v-csum` | |
| **sv** | `bm25` | swedishLight stemmer |
| | `c-bm25` | swedishLight stemmer |
| | `csum` | |
| **uk** | `bm25` | no stemmer & ukStandard stoplist |
| | `c-bm25` | no stemmer & ukStandard stoplist |
| | `csum` | |

## 3 Evaluation

Considering as relevant all documents with either the *Relevant* or *Partially Relevant* judgement, we used a set of binary measures of precision and recall to evaluate the submitted runs.

Moreover, given the lack of official judgements in the first round, we set up a crowdsourcing experiment for five out of eight proposed languages (English, French, German, Italian and Spanish). We gathered assessments for some of the top 100 documents coming from a solid baseline for each language, and we aggregated these assessments using a classic unweighted Majority Vote (MV) algorithm, assigning the *Partially Relevant* label in case of ties.

In this section, we firstly present the evaluation of the submitted runs according to the official judgements (Section 3.1). Then, in Section 3.2 we perform a comparison analysis of our MV pool against the official one, to identify whether a crowdsourced approach can lead to accurate and fast results in this information access scenario aimed at the general public.

### 3.1 Official Evaluation

We evaluated the submitted runs in terms of precision at 5 (p@5), 10 (p@10) and 20 retrieved documents (p@20), average precision (ap) and recall (rec) for seven out of eight proposed languages. Ukrainian runs were not evaluated given the current lack of an official pool. Different runs have been compared using ANalysis Of VAriance (ANOVA), with a significance $\alpha = 0.05$. As post-hoc procedure, we adopted a conservative Tukey's Honestly Significant Difference (HSD).

**Table 5.** Scores achieved by our runs for the **German** language. Marked with $^*$ runs which belong to the same group as the best performing one, according to ANOVA and Tukey's hsd.

| **measure** | **c-bm25** | **v-csum** | **csum** | **bm25** |
|---|---|---|---|---|
| **p@5** | 0.613$^*$ | 0.727$^*$ | 0.627$^*$ | 0.593$^*$ |
| **p@10** | 0.563$^*$ | 0.673$^*$ | 0.570$^*$ | 0.533 |
| **p@20** | 0.525$^*$ | 0.622$^*$ | 0.543$^*$ | 0.525$^*$ |
| **ap** | 0.289 | 0.345$^*$ | 0.307$^*$ | 0.287 |
| **rec** | 0.711$^*$ | 0.708$^*$ | 0.694$^*$ | 0.692$^*$ |

**German** Table 5 reports the scores for our submitted runs on the German language. In terms of precision, overall we observe that `v-csum` is the best performing approach, even though most differences are not statistically significant. A much larger difference can be observed by looking at the average precision: according to this measure, the fusion-based approaches (`v-csum` and `csum`) perform significantly better. In terms or recall, no statistically significant difference can be observed.

**Table 6.** Scores achieved by our runs for the **Greek** language. Marked with * runs which belong to the same group as the best performing one, according to ANOVA and Tukey's hsd.

| **measure** | **c-bm25** | **csum** | **bm25** |
|---|---|---|---|
| **p@5** | 0.753* | 0.627* | 0.593 |
| **p@10** | 0.647* | 0.593* | 0.533 |
| **p@20** | 0.632* | 0.590* | 0.568* |
| **ap** | 0.555* | 0.476* | 0.455* |
| **rec** | 0.962* | 0.871* | 0.861* |

**Greek** Table 6 reports the scores for our submitted runs on the Greek language. Overall, the baseline which employs the conversational formulation (`c-bm25`) is the best method, outperforming the title-only formulation (`bm25`) and the fusion approach (`csum`). The differences in p@5 and p@10 between `c-bm25` and `bm25` are statistically significant, while the remaining score differences are not significant.

**Table 7.** Scores achieved by our runs for the **English** language. Marked with * runs which belong to the same group as the best performing one, according to ANOVA and Tukey's hsd.

| **measure** | **c-bm25** | **v-csum** | **nlex** | **bm25** | **nsle** |
|---|---|---|---|---|---|
| **p@5** | 0.860* | 0.853* | 0.893* | 0.720 | 0.507 |
| **p@10** | 0.827* | 0.823* | 0.900* | 0.690 | 0.513 |
| **p@20** | 0.718* | 0.700* | 0.743* | 0.590 | 0.450 |
| **ap** | 0.277* | 0.300* | 0.306* | 0.227 | 0.159 |
| **rec** | 0.648* | 0.652* | 0.559 | 0.608* | 0.484 |

**English** Table 7 reports the scores for our submitted runs on the English language. In terms of precision, we observe that `nlex` is the best performing approach across the four precision-related measures. `c-bm25` and `v-csum` are also in the best performing group, outperforming `bm25` and `nsle` in a statistically significant way. In terms of recall, on the other hand, the best performing approach is `v-csum`. Both the baseline which employs the conversational formulation of the topic (`c-bm25`) and the two-levels fusion (`v-csum`) are always among the best performing approaches, with no statistical difference from the best method.

**Spanish** Table 8 reports the scores for our submitted runs on the Spanish language. It is interesting to observe that, according to our measures, no fusion run is able to statistically outperform the baselines. Moreover, no single approach outperforms the remaining ones in all measures, not even considering the means. Finally, only `bm25` evaluated with p@20 is significantly worse than the other runs.

**Table 8.** Scores achieved by our runs for the **Spanish** language. Marked with * runs which belong to the same group as the best performing one, according to ANOVA and Tukey's hsd.

| **measure** | **c-bm25** | **v-csum** | **csum** | **bm25** |
|---|---|---|---|---|
| **p@5** | 0.700* | 0.687* | 0.680* | 0.613* |
| **p@10** | 0.693* | 0.713* | 0.620* | 0.580* |
| **p@20** | 0.672* | 0.683* | 0.585* | 0.548 |
| **ap** | 0.165* | 0.170* | 0.172* | 0.146* |
| **rec** | 0.408* | 0.361* | 0.378* | 0.374* |

**Table 9.** Scores achieved by our runs for the **French** language. Marked with * runs which belong to the same group as the best performing one, according to ANOVA and Tukey's hsd.

| **measure** | **c-bm25** | **v-csum** | **csum** | **bm25** |
|---|---|---|---|---|
| **p@5** | 0.800* | 0.740* | 0.727* | 0.667 |
| **p@10** | 0.607* | 0.623* | 0.590* | 0.537* |
| **p@20** | 0.483* | 0.517* | 0.472* | 0.445* |
| **ap** | 0.312* | 0.339* | 0.313* | 0.282 |
| **rec** | 0.763* | 0.657 | 0.691 | 0.693* |

**French** Table 9 reports the scores for our submitted runs on the French language. No approach clearly outperforms the remaining ones, however `c-bm25` is consistently in the top tier.

**Table 10.** Scores achieved by our runs for the **Italian** language. Marked with * runs which belong to the same group as the best performing one, according to ANOVA and Tukey's hsd.

| **measure** | **c-bm25** | **v-csum** | **csum** | **bm25** |
|---|---|---|---|---|
| **p@5** | 0.573* | 0.587* | 0.507* | 0.413 |
| **p@10** | 0.477* | 0.537* | 0.423 | 0.390 |
| **p@20** | 0.358 | 0.465* | 0.405* | 0.325 |
| **ap** | 0.196* | 0.286* | 0.220* | 0.183* |
| **rec** | 0.679 | 0.757* | 0.749* | 0.701 |

**Italian** Table 10 reports the scores for our submitted runs on the Italian language. Fusion-based approaches are consistently in the top performing tier, with the exception of `csum` evaluated with p@10. The title-only baseline `bm25` is consistently outside of the top tier, except for its average precision measure, according to which it is still the worst out of all methods, albeit not significantly.

**Table 11.** Scores achieved by our runs for the **Swedish** language. Marked with $^*$ runs which belong to the same group as the best performing one, according to ANOVA and Tukey's hsd.

| measure | c-bm25 | csum | bm25 |
|---|---:|---:|---:|
| **p@5** | $0.627^*$ | $0.627^*$ | $0.587^*$ |
| **p@10** | $0.580^*$ | $0.610^*$ | $0.563^*$ |
| **p@20** | $0.543^*$ | $0.575^*$ | $0.552^*$ |
| **ap** | $0.504^*$ | $0.460^*$ | 0.418 |
| **rec** | $0.9394^*$ | $0.8294^*$ | $0.8274^*$ |

**Swedish** Table 11 reports the scores for our submitted runs on the Swedish language. The three approaches we considered appear to be equivalent in all metrics, except for `bm25` falling outside of the top tier in average precision.

**Analysis of the factors** We now analyse the interaction between topics, query formulations, systems and their components through the lens of ANOVA. Table 12 contains the p-values and $\omega^2$ computed on the submitted runs using ANOVA. Here $\omega^2$ is the Strength of Association (SOA): it describes how large the effect of a specific component is. For example, a large effect for the topic means that the performance of the system strongly depends on the topic considered. Moreover, if the effect of the system is small, then the score of different systems is similar and there is no strong advantage in using one instead of the other from a performance perspective.

We observe from Table 12 that the effect of the system is always either not significant (meaning that all the systems performed similarly), or significant but negligible. This indicates that, overall, the performance of the systems that produced the submitted runs was comparable. Considering the English language, where we used two pre-trained neural approaches, observing a small effect for the system means that, in this specific setting, there is not a huge advantage in investing into complex architectures.

Moreover, it is interesting to notice that the effect of the topic is almost always small, with few negligible ones and a single medium size effect observed when using average precision as a measure of performance and French as language. This indicates that the topics were almost all equally hard and, considering only the submitted models, the collection appears to be homogeneous in this regard.

Focusing on lexical models, we analyze with ANOVA the importance of different components considered in Section 2.3, considering all the runs summarized in Table 3. Even though most of these runs have not been officially submitted (except for the two baselines for each language), they all contributed to the fusion runs described in Section 2.5. Table 13 reports the p-value and SOA for different factors, namely topic, model, stoplist, stemmer and conversational. The last factor indicates whether we used only the keyword formulation or both the keyword and conversational formulation.

**Table 12.** p-values and Strength of Association (SOA), measured as $\omega^2$, of ANOVA for topic and system factors over the submitted runs considering different languages and measures. *Grey*: non-relevant factors. *White*: Significant yet negligible effects. *Light blue*: Small-size effects. Notice that there is no sizeable effect for either the topics or the systems.

| | | | **de** | **el** | **en** | **es** | **fr** | **it** | **sv** |
|---|---|---|---|---|---|---|---|---|---|
| **p@5** | topic | p-value | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 |
| | | $\omega^2$ | 0.0118 | 0.0123 | 0.0047 | 0.0084 | 0.0174 | 0.0068 | 0.0254 |
| | system | p-value | 0.1420 | 0.0206 | ≤1e-4 | 0.5397 | 0.0704 | 0.0055 | 0.7880 |
| | | $\omega^2$ | — | 0.0008 | 0.0025 | — | — | 0.0007 | — |
| **p@10** | topic | p-value | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 |
| | | $\omega^2$ | 0.0193 | 0.0189 | 0.0063 | 0.0126 | 0.0277 | 0.0125 | 0.0218 |
| | system | p-value | 0.0265 | 0.0567 | ≤1e-4 | 0.0383 | 0.1042 | 0.0017 | 0.7039 |
| | | $\omega^2$ | 0.0005 | — | 0.0032 | 0.0004 | — | 0.0009 | — |
| **p@20** | topic | p-value | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 |
| | | $\omega^2$ | 0.0239 | 0.0225 | 0.0088 | 0.0180 | 0.0368 | 0.0098 | 0.0317 |
| | system | p-value | 0.1143 | 0.2640 | ≤1e-4 | 0.0089 | 0.1355 | 0.0008 | 0.6722 |
| | | $\omega^2$ | — | — | 0.0030 | 0.0006 | — | 0.0011 | — |
| **ap** | topic | p-value | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 |
| | | $\omega^2$ | 0.0504 | 0.0298 | 0.0181 | 0.0243 | 0.0710 | 0.0192 | 0.0436 |
| | system | p-value | 0.0113 | 0.0014 | ≤1e-4 | 0.2692 | 0.0117 | ≤1e-4 | 0.1568 |
| | | $\omega^2$ | 0.0006 | 0.0016 | 0.0034 | — | 0.0006 | 0.0024 | — |
| **rec** | topic | p-value | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 | ≤1e-4 |
| | | $\omega^2$ | 0.0274 | 0.0268 | 0.0362 | 0.0212 | 0.0392 | 0.0487 | 0.0111 |
| | system | p-value | 0.8472 | ≤1e-4 | ≤1e-4 | 0.1534 | 0.0021 | ≤1e-4 | 0.0171 |
| | | $\omega^2$ | — | 0.0026 | 0.0042 | — | 0.0009 | 0.0015 | 0.0008 |

For the current analysis we focus on average precision; similar results obtained with different measures are omitted. Both the topic factor and the model factor exhibit large size effects. A post-hoc analysis, carried out using Tukey's hsd, indicates that lmd, bm25 and dfrinexpb2 are typically the best performing models. Overall, the conversational factor brings about a small yet significant improvement, except for German where keyword-only formulations seem to suffice to express the underlying information needs. Stemmers also have a small yet significant effect size, while the stoplist factor has negligible or non-significant effect.

### 3.2 Pools comparison

We compared our MV pools with the official pools. Table 14 reports the number of judgements for each language in the official and MV pools. As a first step, we analyse the pools agreement on the set of judgements performed on both pools. Given a (topic, document) pair we consider different types of agreement (Figure 3):

- Strong agreement: the relevance judgement for the current pair is equal in the two pools.

**Table 13.** p-values and Strength of Association (SOA), measured as $\omega^2$, of ANOVA on average precision of all lexical runs (Table 3) for the following factors: topic, conversational, stoplist, stemmer and model. *Grey*: non-relevant factors. *White*: Significant yet negligible effects. *Light blue*: small-size effects. *Blue*: medium-size effect. *Dark blue*: large-size effect.

| **factor** | **de** | | **el** | | **en** | | **es** | |
|---|---|---|---|---|---|---|---|---|
| | p-value | $\omega^2$ | p-value | $\omega^2$ | p-value | $\omega^2$ | p-value | $\omega^2$ |
| topic | $\leq$1e-3 | 0.809 | $\leq$1e-3 | 0.738 | $\leq$1e-3 | 0.745 | $\leq$1e-3 | 0.759 |
| conversational | 0.304 | — | $\leq$1e-3 | 0.079 | $\leq$1e-3 | 0.059 | $\leq$1e-3 | 0.025 |
| stoplist | 0.029 | 0.002 | $\leq$1e-3 | 0.010 | $\leq$1e-3 | 0.005 | 0.004 | 0.003 |
| stemmer | $\leq$1e-3 | 0.058 | $\leq$1e-3 | 0.022 | $\leq$1e-3 | 0.019 | $\leq$1e-3 | 0.047 |
| models | $\leq$1e-3 | 0.168 | $\leq$1e-3 | 0.294 | $\leq$1e-3 | 0.166 | $\leq$1e-3 | 0.130 |

| **factor** | **fr** | | **it** | | **sv** | |
|---|---|---|---|---|---|---|
| | p-value | $\omega^2$ | p-value | $\omega^2$ | p-value | $\omega^2$ |
| topic | $\leq$1e-3 | 0.846 | $\leq$1e-3 | 0.680 | $\leq$1e-3 | 0.789 |
| conversational | $\leq$1e-3 | 0.031 | $\leq$1e-3 | 0.007 | $\leq$1e-3 | 0.027 |
| stoplist | 0.037 | 0.001 | 0.525 | — | 0.182 | — |
| stemmer | $\leq$1e-3 | 0.110 | $\leq$1e-3 | 0.032 | $\leq$1e-3 | 0.016 |
| model | $\leq$1e-3 | 0.126 | $\leq$1e-3 | 0.263 | $\leq$1e-3 | 0.157 |

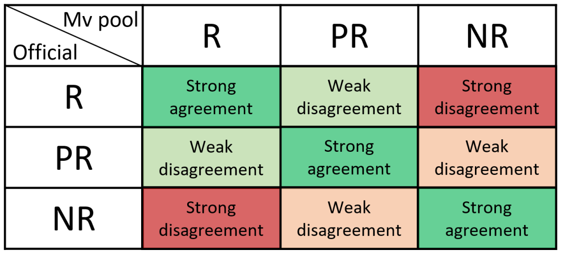
**Table 14.** Number of relevance judgements in official and Majority Vote (MV) pools, common documents in the two pools and % of documents of the official pool also present in MV pool

| | **German** | **English** | **Spanish** | **French** | **Italian** |
|---|---|---|---|---|---|
| **Official Pool** | 5183 | 7242 | 7091 | 4360 | 7680 |
| **Majority Vote Pool** | 975 | 1941 | 1365 | 1119 | 1382 |
| **Common Documents** | 673 | 850 | 761 | 508 | 930 |
| **% overlap** | 0.1298 | 0.1174 | 0.1073 | 0.1165 | 0.1211 |

- Weak agreement: both pools assign a Relevant label, but with different grades of relevance (i.e. PartiallyRelevant in a pool and Relevant in the other, or vice versa).
- Weak Disagreement: one pool assigns a NotRelevant label, the other assigns a PartiallyRelevant label.
- Strong Disagreement: the two pools assign opposite judgements.

Table 15 reports the pool agreement on the common documents. The majority of the judgements for almost all languages Strongly agree with the official pool judgements and Strong disagreements are very rare, indicating a good quality assessment process. There is a non-negligible percentage of Weak disagreements probably due to the lower expertise of the crowd assessors and the possibly different interpretation of the PartiallyRelevant label.

The main weakness of this pool is its small amount of judgements. Table 16 reports the fraction of Relevant and PartiallyRelevant judgements from the



| Mv pool<br>Official | R | PR | NR |
|---|---|---|---|
| R | Strong agreement | Weak disagreement | Strong disagreement |
| PR | Weak disagreement | Strong agreement | Weak disagreement |
| NR | Strong disagreement | Weak disagreement | Strong agreement |

**Fig. 3.** Two judgements for the same (topic,document) pair strongly agree if they are equal, weakly agree if they are both relevant but with different grade, weekly disagree if one is Not Relevant and the other is Partially Relevant, strongly disagree if one is Relevant and one is NotRelevant

**Table 15.** Majority Vote (MV) pools agreement with Official pools, for the 5 available languages

| | German | English | Spanish | French | Italian |
|---|---|---|---|---|---|
| **Strong Agreement** | 0.5750 | 0.5965 | 0.5861 | 0.6102 | 0.4989 |
| **Weak Agreement** | 0.1783 | 0.1788 | 0.2129 | 0.2028 | 0.1387 |
| **Weak Disagreement** | 0.1590 | 0.0859 | 0.1406 | 0.1083 | 0.1968 |
| **Strong Disagreement** | 0.0877 | 0.1388 | 0.0604 | 0.0787 | 0.1656 |

official pool correctly labelled in the MV pool. For example, for English, there are 3276 relevant or partially relevant documents in the official pool. 698 of those documents are also present in the MV pool, 589 correctly labelled and 109 mislabelled, so only 18% of Relevant or PartiallyRelevant judgements in the official pool are present in the MV pool.

**Table 16.** R-PR (Official) Total: total number of Relevant and PartiallyRelevant judgements in the Official pool; R-PR (Official) labelled as P-PR (MV): Relevant or PartiallyRelevant in the official pool that are labelled as Relevant or PartiallyRelevant also in MV; R-PR (Official) labelled as NR (MV): Relevant or PartiallyRelevant in the official pool that are labelled as NotRelevant in MV

| | German | English | Spanish | French | Italian |
|---|---|---|---|---|---|
| **R-PR (Off.) Total** | 2910 | 3276 | 4162 | 2056 | 2673 |
| **R-PR (Off.) labelled as P-PR (MV)** | 344 | 589 | 504 | 337 | 352 |
| **R-PR (Off.) labelled as NR (MV)** | 60 | 109 | 35 | 39 | 59 |
| **% R-PR (Off.) labelled as P-PR (MV)** | 0.1182 | 0.1798 | 0.1211 | 0.1639 | 0.1317 |
| **% R-PR (Off.) labelled as NR (MV)** | 0.0206 | 0.0333 | 0.0084 | 0.0190 | 0.0221 |
| **% R-PR (Off.) not in MV pool** | 0.8612 | 0.7869 | 0.8705 | 0.8171 | 0.8462 |

To inspect the capability of Majority Vote (MV) pool of correctly identifying the best systems, we compute the AP Correlation (APC) between the ranking of the systems induced by the measures computed on the official pool and the ranking of the systems induced by the measures on the MV pool. We considered the whole set of runs for the APC computation (not only the submitted ones). Table 17 shows the AP Correlation values for the available languages and measures. The results show in general a discrete correlation with the official ordering, with English as best language and Italian as worst.

**Table 17.** AP Correlation values between the ranking of the system measures on the official pool and the MV pool

| | **German** | **English** | **Spanish** | **French** | **Italian** |
|---|---|---|---|---|---|
| **p@5** | 0.3990 | 0.5587 | 0.4312 | 0.5414 | 0.4303 |
| **p@10** | 0.3961 | 0.5560 | 0.4582 | 0.4237 | 0.4079 |
| **p@20** | 0.4084 | 0.5777 | 0.4485 | 0.4583 | 0.2494 |
| **ap** | 0.3957 | 0.5851 | 0.3905 | 0.4800 | 0.2928 |
| **rec** | 0.4339 | 0.5584 | 0.5673 | 0.6029 | 0.2195 |

## 4 Conclusions

After the first round, we drew the following preliminary conclusions. In an unsupervised scenario for this novel domain, lexical baselines such as bm25 are confirmed to be reliable and competitive, especially on queries expressed by both keyword and conversational formulations. In the absence of proper domain adaptation, neural models trained on a loosely related domain (medical subset of MS MARCO) are outperformed, however they can still contribute to improve retrieval performance in a rank fusion approach which combines runs from different systems. Availability of query reformulations also improved system performance on average.

Summing up the pool comparison results, the crowdsourced pool and the official pool show a satisfactory relevance judgements agreement, highlighting that crowdsourcing techniques are suitable for the creation of reliable pools in a time-constrained scenario like the one addressed in this pandemic-related initiative.

## References

[1] Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pp. 3613–3618, ACL (2019)

[2] Ferro, N., Harman, D.: CLEF 2009: Grid@CLEF Pilot Track Overview. In: Proc. CLEF (2009)

[3] MacAvaney, S.: Opennir: A complete neural ad-hoc ranking pipeline. In: Proc. of WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020, pp. 845–848, ACM (2020)

[4] MacAvaney, S., Cohan, A., Goharian, N.: SLEDGE: A simple yet effective baseline for coronavirus scientific knowledge search. CoRR **abs/2005.02365** (2020)

[5] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A human generated machine reading comprehension dataset. CoRR **abs/1611.09268** (2016)

[6] Robertson, S.E., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (2009)

[7] Shaw, J., Fox, E.: Combination of multiple searches. In: Proc. of TREC 1994, pp. 105–108 (1994)

[8] 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)

[9] Zhai, C., Lafferty, J.D.: A study of smoothing methods for language models applied to ad hoc information retrieval. SIGIR Forum **51**(2), 268–276 (2017)