

UNIPD IMS group: COVID19- MLIA@Eval

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

Goal:

- 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 a specific topic.

Subtasks:

- Subtask1: high precision
- Subtask2: high recall

Pre-Retrieval

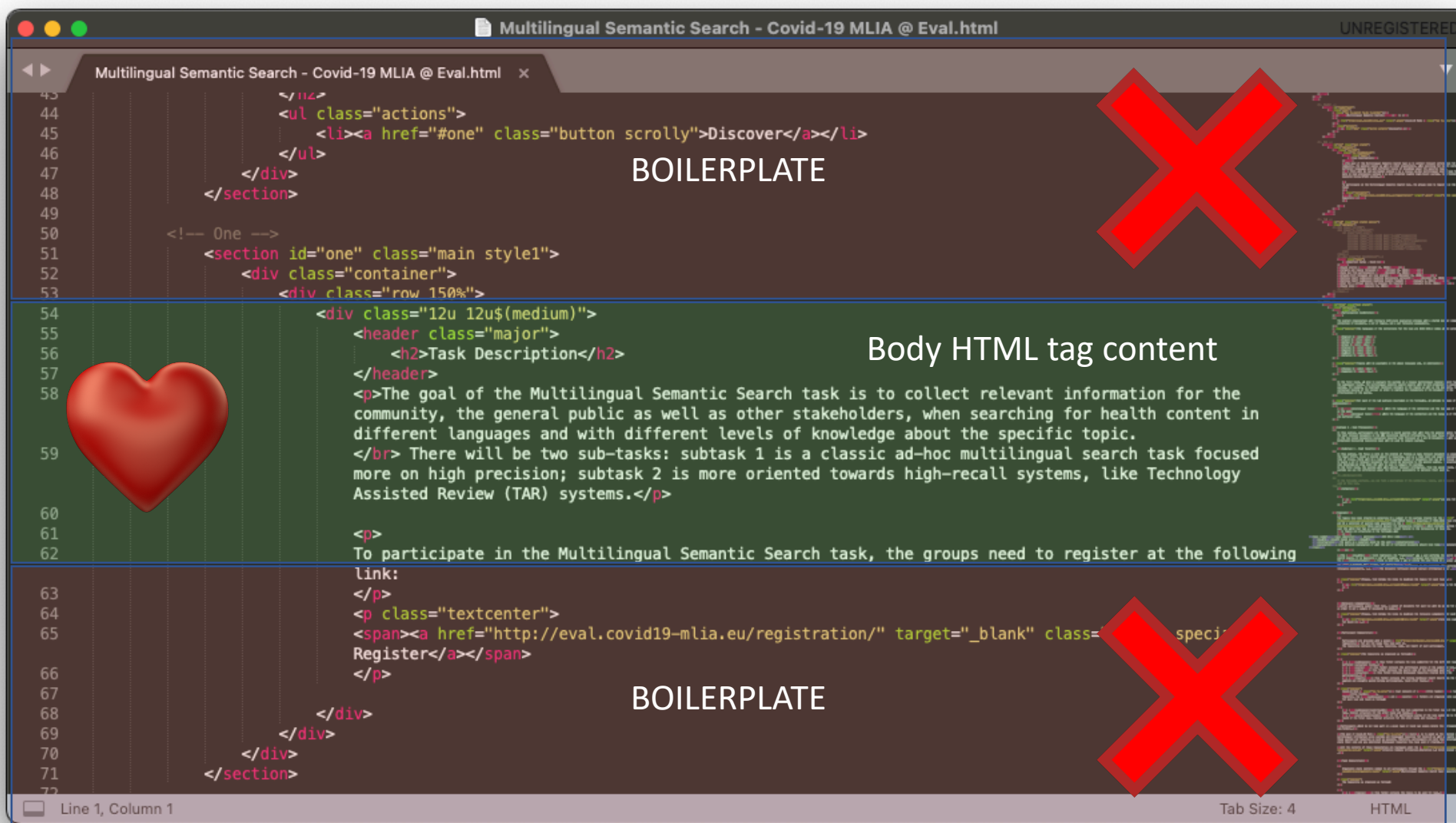
- Doc. Preprocessing
- Query Reformulation

Retrieval

- Lexical
- Neural

Rank Fusion

- CombSUM



```

43     </h2>
44     <ul class="actions">
45       <li><a href="#one" class="button scrolly">Discover</a></li>
46     </ul>
47   </div>
48 </section>
49
50 <!-- One -->
51 <section id="one" class="main style1">
52   <div class="container">
53     <div class="row 150%">
54       <div class="12u 12u$(medium)">
55         <header class="major">
56           <h2>Task Description</h2>
57         </header>
58         <p>The goal of the Multilingual Semantic Search task is to collect relevant information for the
59           community, the general public as well as other stakeholders, when searching for health content in
60           different languages and with different levels of knowledge about the specific topic.
61         </p>
62         <p>There will be two sub-tasks: subtask 1 is a classic ad-hoc multilingual search task focused
63           more on high precision; subtask 2 is more oriented towards high-recall systems, like Technology
64           Assisted Review (TAR) systems.</p>
65
66         <p>
67           To participate in the Multilingual Semantic Search task, the groups need to register at the following
68           link:
69           <p class="textcenter">
70             <span><a href="http://eval.covid19-mlia.eu/registration/" target="_blank" class="speci
71             Register</a></span>
72           </p>
73         </div>
74       </div>
75     </div>
76   </section>

```

Line 1, Column 1

Tab Size: 4

HTML

- We collaborated with 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;
- We first asked the students to provide a few **terminological variations** to groups of one or more medical terms in each topic keywords field;
- We then **automatically generated the query variations** by replacing all combinations of these terms in the topic keyword field.

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

- We performed document retrieval using all the combinations of the following stoplists, stemmers and retrieval models for each language:

Lang.	Stoplists	Stemmers	Ranking Fun.	GoP size
de	bbalet*, ranksnl*, gh*, lucene, nostop	nostem, german, germanLight		75
el	nostop, bbalet*, ranksnl*, gh*, lucene	nostem, greek	bm25 tf-idf	50
en	nostop, lucene, indri [†] , atire [†] , okapi [†]	nostem, porter, lovins	lmd lmjm	75
es	nostop, bbalet*, ranksnl*, gh*, lucene	nostem, spanishLight, snowball	dfrinexpb2	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

Table 1: 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.

- We also reranked the top 500 documents of each topic of a run computed with Anserini BM25 (using the default stemmer and stoplist options) using the SLEDGE¹ framework, a recent neural approach for reranking relying on SciBERT.
- We used two different pre-trained models available online, one trained on the MS-MARCO dataset and the other trained on only the subset of medical documents in it¹.

¹ *The models are described and downloadable from: [//github.com/Georgetown-IR-Lab/covid-neural-ir](https://github.com/Georgetown-IR-Lab/covid-neural-ir).*

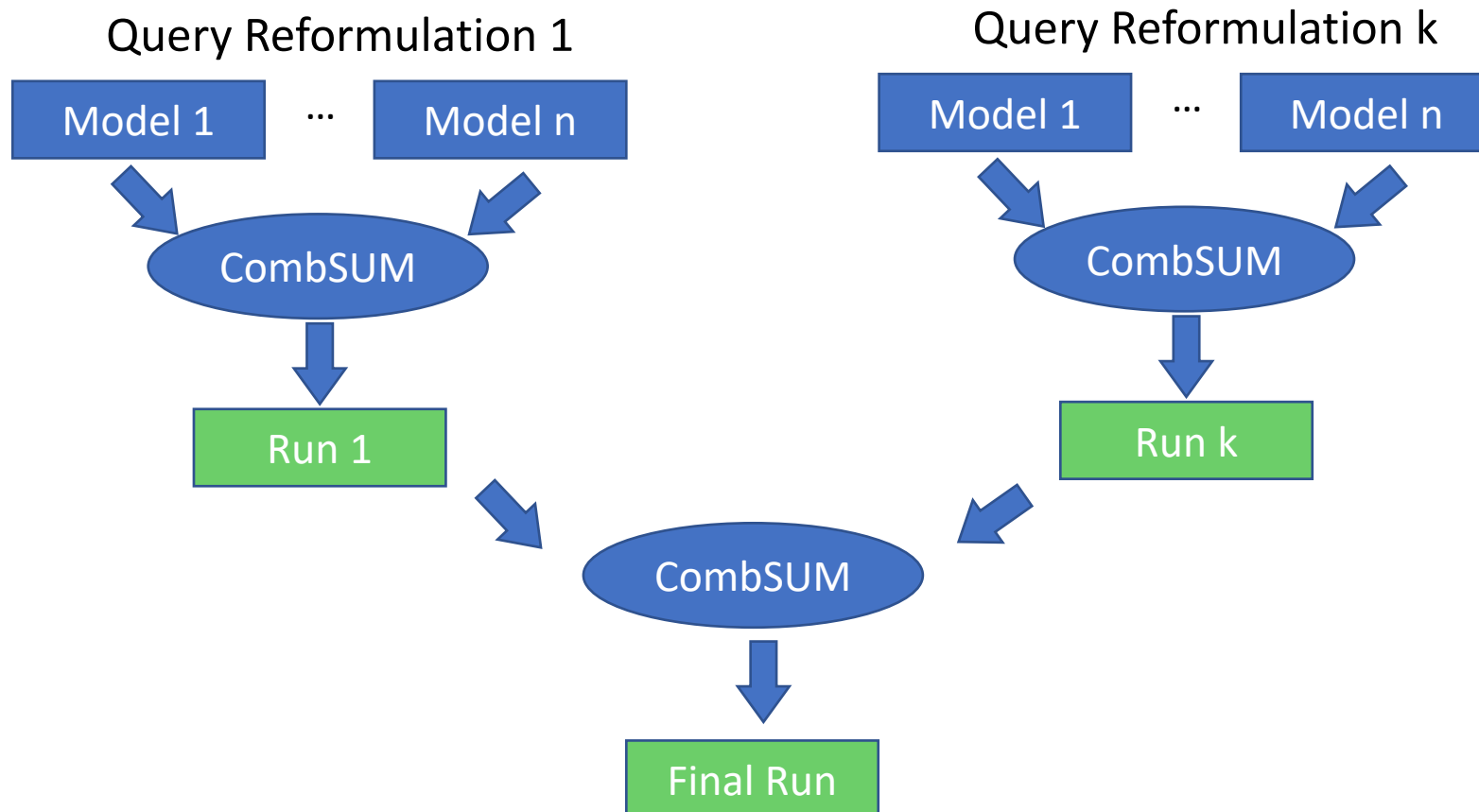


Table 1: AP achieved in our submitted runs for different languages.

Lang	c-bm25	v-csum	csum	bm25	nsle	nlex
DE	0.289	0.345	0.307	0.287	-	-
EL	0.555	-	0.476	0.455	-	-
EN	0.277	0.300	-	0.227	0.159	0.306
ES	0.165	0.170	0.172	0.146	-	-
FR	0.312	0.339	0.313	0.282	-	-
IT	0.196	0.286	0.220	0.183	-	-
SV	0.504	-	0.460	0.418	-	-

- bm25: BM25 with default Lucene parameters (keyword only formulation of each topic);
- c-bm25: same as above, with queries combining both keyword and conversational formulations of topics;
- csum: one-stage fusion of all the lexical runs, using only the keyword-only formulation of a query;
- v-csum: two-stage fusion, using all the available topic formulations and lexical runs (DE, EN, ES, FR, IT only);
- nlex: three-stage fusion, using all the available topic formulations, lexicalruns and neural runs (EN only);
- nsle: the output of sledge-med (EN only).

Mv pool \ 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

Table 1: Majority Vote 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

- When **no relevance judgements** are available, **BM25** is the most reliable solution as a single system
- **Rank fusion** of multiple lexical models can give a **cheap and consistent performance boost** across multiple languages
- When possible, relying on human annotators or on an ontology to perform query **reformulation can further boost performance** in a noticeable and reliable way
- **Neural models alone**, even when trained on a same-domain corpus **cannot achieve competitive performances** but become very **useful when combined through rank fusion** techniques to lexical models

Thank you!

Questions?