Covid-19 MLIA Information Extraction Task Round 1 Presentation and Main Findings

Cyril Grouin[1], Thierry Declerck[2], and Pierre Zweigenbaum[1]

[1] Université Paris-Saclay, CNRS, LIMSI, France
{cyril.grouin,pz}@limsi.fr
[2] DFKI, Germany
thierry.declerck@dfki.de

Abstract. In this paper, we present the information extraction task proposed in the first round of the Covid-19 MLIA @ Eval Initiative. During this first round, we proposed to identify six categories of information potentially relevant for the Covid-19 issue (sign or symptom and disease, medical test, drug and treatment, legal rules, everyday life actions, and findings), on texts available in seven languages (English, French, German, Greek, Italian, Spanish, and Swedish). Since no gold standard annotations were given, the participants reused their existing tools and ressources. Four teams participated in this task.

1 Introduction

The goal of the Information Extraction task is to identify medical information in texts. We defined six major types of entities to be identified. Those categories are mainly related to the Covid-19 issue. The main objective is to mine texts in order to access relevant information concerning the Covid-19, and more specifically information that may help the health professional to find outcomes.

The first round of this task started on October 23rd. During this round, participants will only have access to unannotated data in a plain text format. The evaluation will consist in a ROVER of system outputs [1]. We encourage the participants to try experimental methods and to submit several system outputs in order to exchange different views during the discussion at the virtual meeting. The submissions were due on November 27th.

Thirty-two teams registered from seventeen countries (Australia, Botswana, Canada, China, France, India, Italy, Mexico, Pakistan, Portugal, Saudi Arabia, Spain, Switzerland, Tunisia, Turkey, The United Kingdom, The United States of America). Almost all participants belong to a university.

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

2 Task and Corpora

2.1 Task description

In this task, we focused on six categories of information related to the Covid-19 issue:

- **drug names, treatments, general intervention:** this category concerns both commercial and generic names of drugs, as well as general intervention in the health domain; elements from this category usually come from advices from a professional (medical doctor, pharmacist) or from self-medication, e.g., Posaconazole AHCL, Allegra, Fexofenadine HCL, Xarelto, quarantine
- **signs, symptoms, diseases:** this category deals with medical problems and merges together all signs, symptoms, and diseases shortness of breath, extreme fatigue, fever, skin infection, weightloss
- **findings, efficacy of treatments:** this category is more complex since it concerns all elements related to positive or negative effets of treatments, including non expected stuff
- **tests:** this category concerns all tests performed to diagnose medical problems such as blood sample, physical exam, serological test
- **behaviors, everyday life actions:** this category concerns all actions performed by each of us such as to wash one's hands, to cough into his elbow, to self-confine, use of face masks, physical distancing
- **legal dispositions, regulations:** this category concerns all actions decided by local or national authorities (Government, Ministry, etc.), such as to download the employer certificate, list of authorized move, prolonged border closure, closure of educational institutions

All system outputs are expected to be in the BRAT [3] annotation format (i.e., a tabular *.ann file for each *.txt file, composed of three columns: (i) an annotation ID, (ii) category, starting offset, ending offset, and (iii) the corresponding text span). An example is shown below:

```
T1      drug-trt 34 68      Irbesartan Hydrochlorothiazide BMS
T2      sosy-dis 116 125    dizziness
```

Since the evaluation takes into account the position of the extracted spans in the texts, the participants must carefully check the offsets they compute.

2.2 Corpora

The corpora used in the information extraction task are exactly the same than those from the machine translation task. We decided to propose the same content in order to allow the reuse of findings from task 1 in the two other tasks. Nevertheless, this choice implies that sentences extracted from the parallel sentences do not refer anymore to a coherent document since pairs of sentences from the machine translation task do not needs to keep the whole content of a single document. This also implies that content is parallel in each language, which can be helpful for participants to produce multilingual tools.

Corpora are available in seven languages: English, French, German, Greek, Italian, Spanish, and Swedish. For each language, we proposed up to 12 files of content. Each file is composed of one sentence per line. As previously explained, sentences are independant between them and do not compose a consistent content. Table 1 presents the number of files available in each language from the training and test datasets, as well as the total number of words and sentences to process. For the test dataset, we decided to split the files into smaller ones (for a maximum number of 2 500 sentences per file) in order to make it easier both the computation of offsets of characters for each annotated span and the output evaluation; this splitting process explains the apparent increase number of files between train and test.

		English	French	German	Greek	Italian	Spanish	Swedish
Train	Files	12	12	12	10	12	12	9
	Words	19410k	22579k	16129k	16352k	18188k	22260k	13163k
	Sentences	1004k	1004k	926k	834k	900k	1028k	806k
Test	Files	52	52	18	5	7	32	12
	Words	1768k	2141k	198k	55k	108k	1165k	129k
	Sentences	98k	98k	11k	2830	5338	55k	9062

Table 1. Number of files, words and sentences ('k' stands for kilo: 19410k means 19,410,000 words) per language from the training and test datasets

Table 2 gives a few sentences from the file 2730.txt in four languages (English, French, German, and Spanish). Note that some sentences are parallel in several languages.

3 Results

3.1 Submissions

We received eleven submissions from four teams: two compagnies, Accenture (USA) and Innoradiant Research Group (France); and two academic teams, SWLab/University of Cagliari (Italy) and ZHAW/School of Engineering (Switzerland). Despite gentle reminders to the other participants, we did not receive any additional submissions. We present in the table 3 the number of submissions per language for each participant.

While we received several submissions for English, we did not received any submission for two languages (French and Swedish), and we received submissions from only one team for the three other languages (German, Greek, and Italian).

Among all received submissions, we observed a small formatting error for one team (use of tabulations instead of spaces between label, beginning offset and ending offset) that we automatically correct in order to perform the evaluation.

English	More factsheets
	For more information check the ECDC website: https://www.ecdc.europa.eu/en/measles/facts/factsheet
	on the third to seventh day, the temperature may reach up to 41 °C;
	Two doses of the vaccine are needed for maximum protection.
	The only protection against measles is vaccination.
French	Autres fiches d'information
	Pour en savoir plus, consultez le site web de l'ECDC: https://www.ecdc.europa.eu/en/measles/facts/factsheet
	du troisième au septième jour, la température peut atteindre 41 °C;
	Deux doses du vaccin sont nécessaires pour obtenir une protection maximale.
	La seule protection contre la rougeole est la vaccination.
German	Bei dieser Person besteht auch das Risiko von Komplikationen.
	Dies wird das Antigen genannt.
	Immunität hält in der Regel über mehrere Jahre an, mitunter ein ganzes Leben lang.
	Wie Impfstoffe wirken
	Diese ,,Gemeinschaftsimmunität" kann nur funktionieren, wenn genügend Personen geimpft sind.
Spanish	Para obtener más información, visite el sitio web del ECDC: https://www.ecdc.europa.eu/en/measles/facts/factsheet .
	entre el tercer y el séptimo día, la temperatura puede llegar hasta 41 °C;
	El 30 % de los niños y de los adultos infectados con sarampión pueden desarrollar complicaciones.
	La vacuna TV es segura y efectiva y tiene muy pocos efectos secundarios.
	La primera dosis se administra entre los 10 y 18 meses de edad en los países europeos.

Table 2. Extract from the file 2730.txt in a few languages

3.2 Evaluation

Since there is no ground truth for this first round, we planned to evaluate the system outputs based on a ROVER computed on the outputs provided by all participants, following the general method designed by Fiscus [1]. As done by Rebholz-Schuhmann et al. [2], we worked at the character level to align all outputs and to compute the majority vote.

Results are evaluated using the traditional metrics used in information extraction: precision, recall, and F-measure. Those metrics are used to compute scores for each entity class. Nevertheless, since this first round is explotary, we chose to mainly evaluate the system outputs using a precision, which allows us to evaluate how a system performs well among all provided predictions.

English For English, due to the low number of submissions and so as to check whether a ROVER-based evaluation is still relevant, we proposed two evaluations: first, using a ROVER, and second, using gold standard annotations.

Team	DE German	EL Greek	EN English	ES Spanish	FR French	IT Italian	SV Swedish
Accenture	0	0	1	0	0	0	0
Innoradiant	0	0	2	0	0	0	0
SWLab	0	0	1	0	0	2	0
ZHAW	1	1	1	1	0	0	0

Table 3. Number of submissions per language for each participant

ROVER As initially planned, we produced a ROVER based on all submissions. In order to do not favour the participants that submitted several runs, which would potentially increase the weight of their predictions in the produced reference, we first concatened all submissions from each participant into a single submission file. We then produced a ROVER based on four outputs, corresponding to the four participants. We kept annotations when they were shared by at least two participants (we present in table 4 a short sample of output alignements and the annotation kept in the final column for the sequence *"the spread of COVID-19."* found in the file 3693-aa.txt).

Offset	Character	Innoradiant	SWLab	ZHAW	Accenture	ROVER
849	t	O	O	O	O	O
850	h	O	O	O	O	O
851	e	O	O	O	O	O
852	SPACE	O	O	O	O	O
853	s	O	O	B-findings	O	O
854	p	O	O	I-findings	O	O
855	r	O	O	I-findings	O	O
856	e	O	O	I-findings	O	O
857	a	O	O	I-findings	O	O
858	d	O	O	I-findings	O	O
859	SPACE	O	O	O	O	O
860	o	O	O	O	O	O
861	f	O	O	O	O	O
862	SPACE	O	O	O	O	O
863	C	B-sosy-dis	O	B-sosy-dis	O	B-sosy-dis
864	O	I-sosy-dis	O	I-sosy-dis	O	I-sosy-dis
865	V	I-sosy-dis	O	I-sosy-dis	O	I-sosy-dis
866	I	I-sosy-dis	O	I-sosy-dis	O	I-sosy-dis
867	D	I-sosy-dis	O	I-sosy-dis	O	I-sosy-dis
868	-	I-sosy-dis	O	I-sosy-dis	O	I-sosy-dis
869	1	I-sosy-dis	O	I-sosy-dis	O	I-sosy-dis
870	9	I-sosy-dis	O	I-sosy-dis	O	I-sosy-dis
871	.	O	O	O	O	O

Table 4. Sample of output produced by the ROVER: offset, character, predictions from each team (Innoradiant, SWLab, ZHAW, Accenture), and final annotation kept

Table 5 presents the total number of predictions for each team, after concatenation of predictions from all submissions, and the results for each category and globally.

Team	Predictions	behav.	drugs	find.	legal	sosy	tests	overall
Innoradiant	52 992	1.000	.976	.000	.000	.995	.977	.990
ZHAW	57 061	1.000	.963	.000	1.000	.988	.919	.982
Accenture	7 010	.000	.076	.000	1.000	.022	.000	.034
SWLab	170	.000	.000	.000	.000	.004	.105	.004

Table 5. Number of predictions and results (precision) for each category (behavior, drugs/treatments, findings, legal rules, sign or symptoms/diseases, tests) and globally (overall) for each submission, using the ROVER (52 files)

Gold standard annotations We manually produced gold standard annotations for a selection of 9 files from the test dataset, chosing the most frequently annotated files by the four participants, assuming those files are the most relevant in the test dataset. Since a human annotation process was not planned, only one annotator participated in this work (no inter-annotator agreement may be computed), for a total number of 1740 annotations: 1173 signs, symptoms and diseases,[3] 228 behavior and everyday life actions, 160 legal rules, 132 drugs and treatments, 46 medical tests, and only 1 findings.[4]

We performed the evaluation only on those nine files using the BRATeval tool, using a relax evaluation mode, which allows for distance errors of one character w.r.t. the offsets from the reference. Table 6 presents the total number of predictions on those files and the results for each submission.

German, Greek, and Italian For the three other languages, since we only received submissions from one team for each language (SWLab for Italian, ZHAW for German and Greek), we can not produce any ROVER and we can not perform any large-scale evaluation for those submissions.

4 Discussion

As shown in table 4, the ROVER is based on the predictions mostly produced by the participants. One main default of this method is that a participant may

[3] Among all signs, symptoms, and diseases identified in the test dataset, we annotated 588 occurrences (50.1%) of *COVID-19, Covid-19, covid19, COVID19, Coronavirus Disease 2019* forms, and 199 occurrences (17.0%) of *coronavirus, Coronavirus*. Other annotations mainly concern *fever, cough, anxiety, worried, GAD* (General Anxiety Disorder), etc.

[4] We identified the following phrase: *"only about 2% of the population has developed antibodies"*.

Team	Predictions	behavior (n=228)	drugs (132)	find. (1)	legal (160)	sosy (1173)	tests (46)	overall
Innoradiant	3893	.447	.197	.000	.000	.720	.196	.564
ZHAW	3796	.088	.189	.000	.031	.398	.304	.305
Innoradiant (long)	263	.355	.000	.000	.000	.000	.000	.047
Accenture	559	.000	.083	.000	.000	.008	.000	.012
SWLab, run #2	9	.000	.000	.000	.000	.001	.000	.001
SWLab, run #1	8	.000	.000	.000	.000	.000	.000	.000
SWLab, run #3	9	.000	.000	.000	.000	.000	.000	.000

Table 6. Number of predictions and results (precision) for each category (behavior, drugs/treatments, findings, legal rules, sign or symptoms/diseases, tests) and globally (overall) for each submission, using the gold standard annotations (9 files). The number of annotations per category in the reference is presented between parentheses

have produced a relevant annotation, but this annotation will not be kept in the ROVER output since the other participants did not produced a prediction for this span, or they used a distinct label. This has two consequences: first, we miss this relevant prediction, which is a pity for an exploratory information extraction task; second, the evaluation process will consider this prediction as a false positive, and thus, it decreases the precision value.

Nevertheless, a comparison between tables 5 and 6 shows that there is no difference in the final ranking of the participants. Obviously, the observed difference of number of predictions between each participant (a lot of predictions for Innoradiant and ZHAW, a moderate number of predictions for Accenture, and a very low number of predictions for SWLab) is also an explanation for the similar ranking achieved using the two evaluations.

In this task, we considered all predictions as being relevant in their category if shared by several participants (ROVER) or common with the gold standard annotation. While this information extraction task mainly focused on the Covid-19 issue, one may argue the need for a slightly better adapted predictions and annotations to this issue. As an example, if the symptom *anxiety* is found in a text, but this anxiety is not related to the Covid-19, identifying and extracting this information is no longer relevant for the general Covid-19 general purpose. This implies a better understanding of the texts to process and it makes more complex the task itself.

References

[1] Fiscus, J.G.: A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In: IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings, pp. 347–54 (1997)

[2] Rebholz-Schuhmann, D., Yepes, A.J., Li, C., Kafkas, S., Lewin, I., Kang, N., Corbett, P., Milward, D., Buyko, E., Beisswanger, E., Hornbostel, K.,

Kouznetsov, A., Witte, R., Laurila, J.B., Baker, C.J., Kuo, C.J., Clematide, S., Rinaldi, F., Farkas, R., Móra, G., Hara, K., Furlong, L.I., Rautschka, M., Neves, M.L., Pascual-Montano, A., Wei, Q., Collier, N., Chowdhury, M.M., Lavelli, A., Berlanga, R., Morante, R., Van Asch, V., Daelemans, W., Marina, J.L., van Mulligen, E., Kors, J., Hahn, U.: Assessment of NER solutions against the first and second CALBC silver standard corpus. J Biomed Semantics **2**(Suppl 5) (2011)

[3] Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: BRAT: a web-based tool for NLP-assisted text annotation. In: Proc of EACL Demonstrations, pp. 102–7, ACL, Avignon, France (2012)