Covid-19 MLIA @ Eval

Information Extraction Task
Round 1 Presentation and Main Findings

Cyril GROUIN (Université Paris-Saclay, CNRS, LISN, France)
Thierry DECLERCK (DFKI, Germany)
Pierre ZWEIGENBAUM (Université Paris-Saclay, CNRS, LISN, France)

Introduction

- Objective: to identify relevant medical information in texts related to the Covid-19 issue
- Organization: first round
  - October 23rd: access to unannotated training data
    - 32 registrations from 17 countries
  - November 27th: submissions due

Task Description

Six categories of information to be found:

- **Drug names, treatments, general interventions:** *Posaconazole AHCL, Allegra, Fexofenadine HCL, Xarelto, Quarantine*
- **Signs, symptoms, diseases:** *shortness of breath, extreme fatigue, fever, skin infection, weightloss*
- **Findings, efficacy of treatments:** positive or negative effects, unexpected stuff
- **Tests:** *blood sample, serological test*
- **Behaviors, everyday life actions:** *to wash one's hands, to cough into his elbow, to self-confine, use of facial masks, physical distancing*
- **Legal dispositions, regulations from local/national authorities:** *employer certificate, list of authorized move, prolonged border closure*

Corpora

- Same datasets as Machine Translation task

	English	French	German	Greek	Italian	Spanish	Swedish
Files (train)	12	12	12	10	12	12	9
Sent. (train)	1004k	1004k	926k	834k	900k	1028k	806k
Files (test)	52	52	18	5	7	32	12
Sent. (test)	98k	98k	11k	2830	5338	55k	9062

Corpora

- All sentences have been extracted from the MT task TMX files
  - sentences are not related together: the sequence of text is lost
  - no annotation available
    - allows a participant to find their own content based on general definitions
    - does not allow to train systems based on existing annotations
    - evaluation does not rely on a gold standard
- Test dataset: all files composed of at most 2500 sentences (more files but less content)

Corpora

Sample sentences:

- on the third to seventh day, the temperature may reach up to 41 °C;
- Two doses of vaccine are needed for maximum protection
- entre el tercer y el séptimo dia, la temperatura puede llegar hasta 41 °C;
- Se trata de una enfermedad grave que puede causar complicaciones e incluso la muerte
- Il 30 % dei bambini e degli adulti infettati dal morbillo può sviluppare complicanze, che possono includere infezioni alle orecchie e diarrea.
- Tavolta i pazienti sviluppano complicanze batteriche a seguito di un’infezione influenzale e devono essere sottoposti a terapia antibiotica.

Submissions

4 participants:

Team	Status	Languages
Accenture	Company (USA)	English (1 run)
Innoradiant	Company (France)	English (2 runs)
SWLab	Academic (Italy)	English (1 run), Italian (2 runs)
ZHAW	Academic (Switzerland)	German (1 run), Greek (1 run), English (1 run), Spanish (1 run)

ROVER at character level

Offset	Character	Team #1	Team #2	Team #3	Team #4	ROVER
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

English Evaluation #1

ROVER produced on a combination of all submissions made by the participants

- 4 participants = 4 outputs (including participants that submitted several runs): does not give more weight to predictions made in several runs
- for each team, among all runs, the majority prediction made in all runs is kept in the combined version
- annotations kept if shared by at least 2 participants

English Evaluation #2

Manual gold standard annotations

- 9 files (the most annotated by all participants)
- only 1 annotator; no inter-annotator agreement
- 1740 manual annotations:
  - 1173 signs, symptoms, diseases; 228 behavior everyday life actions; 160 legal rules; 132 drugs treatments general interventions; 46 medical tests; 1 findings

Evaluation in Other Languages

- Impossible to produce a ROVER:
  - German: 1 run from one team (ZHAW)
  - Greek: 1 run from one team (ZHAW)
  - Italian: 2 runs from one team (SWLab)
- Nothing to evalute:
  - French & Swedish: 0 submission

ROVER Results (precision)

Team	Predict numb.	Behav.	Drugs	Find.	Legal	Sosy	Tests	Overall
Innoradiant	52,992	1.000	0.976	0.000	0.000	0.995	0.977	0.990
ZHAW	57,061	1.000	0.963	0.000	1.000	0.988	0.919	0.982
Accenture	7,010	0.000	0.076	0.000	1.000	0.022	0.000	0.034
SWLab	170	0.000	0.000	0.000	0.000	0.004	0.105	0.004

Gold Standard Results (Precision)

Team	Pred. nb	Behav. (n=228)	Drugs (132)	Find. (1)	Legal (160)	Sosy (1173)	Tests (46)	Overall (1740)
Innorad	3893	0.447	0.197	0.000	0.000	0.720	0.196	0.564
ZHAW	3796	0.088	0.189	0.000	0.031	0.398	0.304	0.305
Inno. (long)	263	0.355	0.000	0.000	0.000	0.000	0.000	0.047
Accent	559	0.000	0.083	0.000	0.000	0.008	0.000	0.012
SWL#2	9	0.000	0.000	0.000	0.000	0.001	0.000	0.001
SWL#1	8	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SWL#3	9	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Methods

- POS taggers: to detect boundaries
- BERT models: to annotate entities
- Okgraph library (Univ. Cagliari): word embeddings and unsupervised algorithms
- cTAKES NER + the UMLS Terminology Service

Discussion

- Round #1: no annotations were available
  - ✖ no training data for systems: many teams registered but did not participate due to this lack of annotations
  - ✓✖ no gold standard for evaluations: OK if several participations in each langage (English only)
- ✓ Results are similar between ROVER and gold standard evaluations
- ✓ Similar understanding of what kind of information to annotate across participants
- Round #2: new data w/ context; gold standard annotations expected for test

Conclusion

Thank you for your participation in this 1st round!