INNORADIANT
Voice of the User

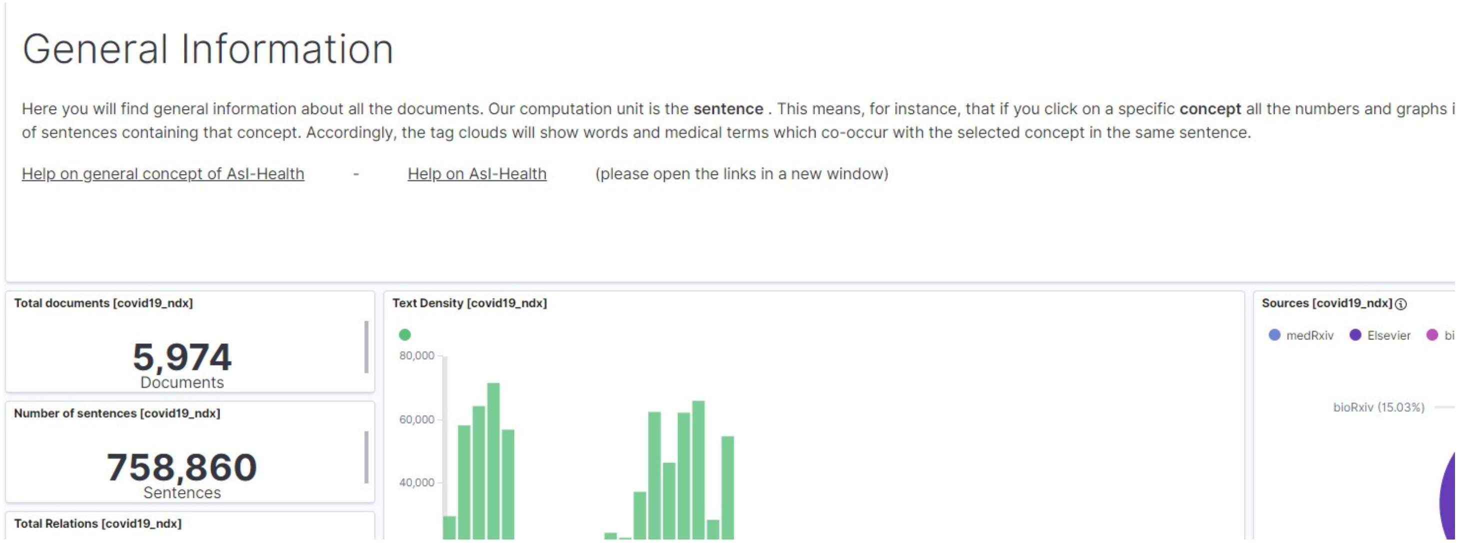
Fine Tuning Named Entity Extraction

(Paolo Curtoni, Luca Dini)

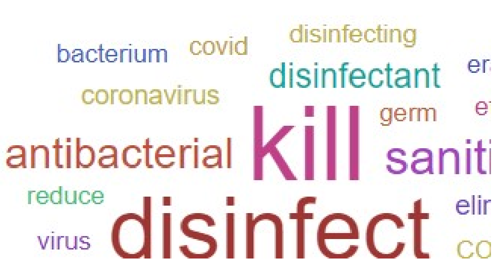
Covid-19 MLIA @ Eval

Motivations

- A new playground re-using previous COVID-19 related knowledge (http://semarillion.com/app/kibana#/dashboard/cf6b6080-9440-11ea-8a42-7bbefcc29248):



- Gaining information on usages in the COVID-19 period.



Approches

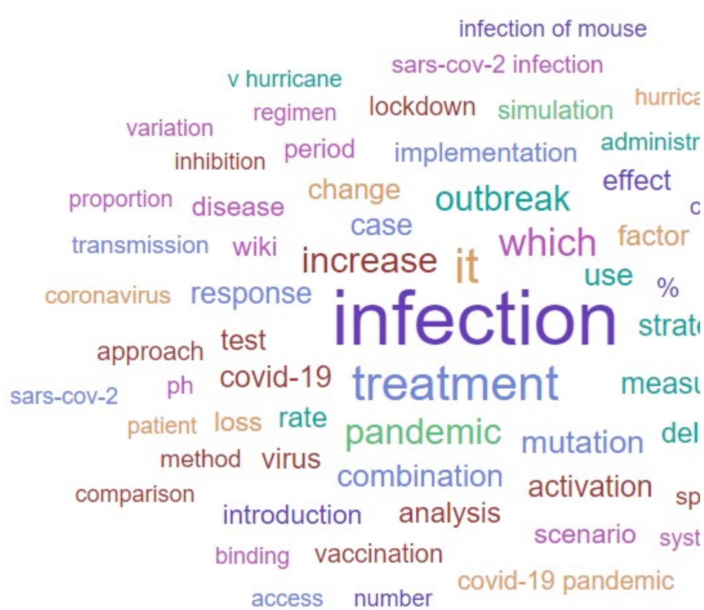
- Accurate, grammar-based identification of behaviors associated to Covid-19 pandemic:
    - Inspired by Semarillion/COVID-19
    - Based on dependency parsing and semantic rules
    - High precision, low recall
- Massive identification of medical named entities:
    - Based on an existing system
    - Semi supervised filtering
    - Classification approach.

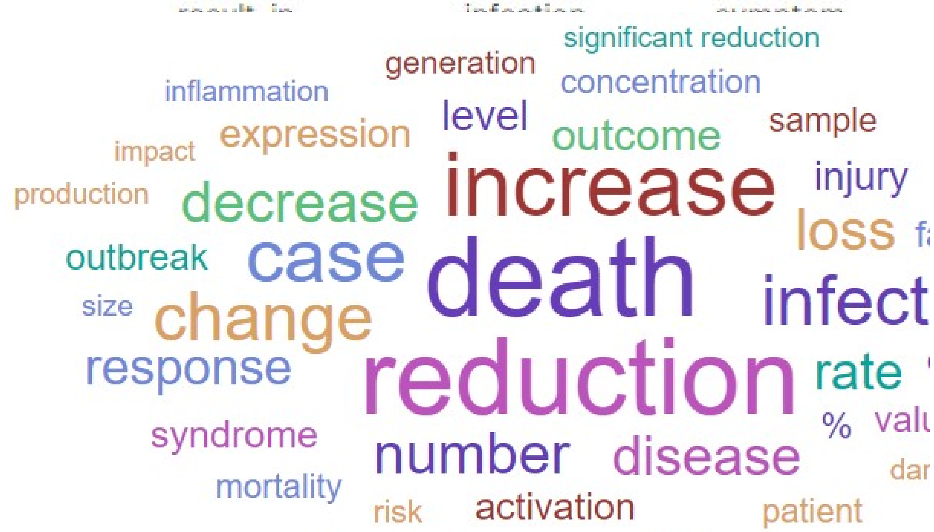
Grammar based behavior recognition (1)

- The input is represented by a set of triples encoding all relations (subject, predicate, object) detected in all corpora.
    - Triples are obtained via a semantic oriented dependency parser coupled with manual rules of semantic simplification (Semarillion). Ex.:
        - Featurization of negation and prepositions.
        - Passive undoing.
        - Modality

Relation Main Table [covid19_ndx]

Predicate	Subject	Object
result_in	infection	disease
result_in	infection	syndrome
result_in	infection	activation





Grammar based behavior recognition (2)

- A second module contains manually written rules performing graph labelling:
  - Each sentence is a graph where node identity is either lemma determined or WikiData id
  - Rules have the form:

```
{
  "id" :  "wash_clean",
  "preds" : [ "wash", "clean", "disinfect" ,"soap","sa
  "objs" : [ "hand", "surface" ]
}, {
  "id" :  "use_wear",
  "preds" : [ "use" ,"wear"],
```

  - Graph matching was enhanced with W2V matching with different thresholds

Grammar based behavior recognition (3)

- Results so far are extremely deceptive with a precision of 0.35.
- Possible reasons:
  - For many texts we could not obtain a reasonable dependency representation.
  - At configuration phase there was the assumption that the corpus contained "emotional behavior", i.e. subjective reactions.
  - Screwed offsets?
  - Human introspection without any gold standard.

Massive identification of medical named entities (1)

- Use as a basis an existing medical named extraction sys-tem, namely Apache cTAKES (https://ctakes.apache.org/):
  - Default configuration
  - No re-training
  - Access to UMLS Terminology Services (https://uts.nlm.nih.gov/)
- Problem: Massive over-generation of medical terms and low recall for “common” terms.
- Measures to face the problem:
  - Sentence classification
  - Terminology filtering
  - Seeded expansion

Massive identification of medical named entities (2)

- Problem: we had the strong feeling that the same parameters could not account for medical and non medical texts.
- Solution: classify sentences according to med/no med categories:
  - Create a gold standard randomly sampling from “EU Press Corner”, “EUR-Lex” and MEDISYS (Multilingual Search Track).
  - Train a simple BERT classifier (bert-base-cased) to make the difference between medical and non medical.
  - Filter cTakes predictions on the basis of probability and category.

Massive identification of medical named entities (2)

- A further refinement has been performed by using a terminology induced on the corpus.
- The selected system is TermSuite (Cram & Daille 2016)
- Any cTakes term not appearing also in the terminology was discarded
- As for increasing the recall (behaviours and non-jargon NE):
  - Manually select for each category ten best matching Terms/NE
  - Match, via W2V other terms whose similarity to manually selected was above a certain threshold
  - Generalize the selected expression via predefined patterns and applies them to the corpus (e.g. Noun-prep-Noun -> Noun-prep-(det|adj)*-Noun)

Future Work

- In many parts of the whole system there are thresholds (similarity, acceptability, etc.) which need to be set. The presence of a training set will allow a better tuning.
- For Grammar based:
  - use the training set to induce rule by adopting relation extraction and Knowledge graph completion techniques.
- For massive NE extraction:
  - Perform threshold tuning.
  - Test different proximity algorithms
- General:
  - why not something slightly less NER oriented for behaviours?
  - Creation of a Social Network corpus?

THANKS