INNORADIANT
Voice of the User

AsI-Health General Principles

This document was produced in the context of Innoradiant's effort to provide people engaged on Covid-19 research a text mining platform to explore relevant scientific papers. In this document we shortly describe the general concepts underlying AsI-Health i.e. relations, subjects, objects and nominals.

Contents

1 General Principles

1.1 Relations

AsI-Health is a platform for disclosing semantic information captured from a corpus of scientific paper and preprint about Covid-19. The core of the semantic information is represented by **relations**, also called **triples**. Shortly a relation is just a fact, a claim, an event which relates two actors/entities. For instance in

> *Comorbidities significantly increased the death risk.*

There is a relation between the entity *comorbidity* (the **subject**) and the entity *risk* (the **object**) which is mediated by the verb *increase* (the **predicate**). We informally represent such a relation as `comorbidity-increase-risk`.

Besides capturing standard direct syntactic relations, AsI-Health uses Natural Language Processing techniques to **capture relations that might be hidden by syntax**. For instance the relation `comorbidity-increase-risk` could have been extracted from

> *The risk was increased by comorbidities.*

or

> *Comorbidities, which increase the risk, where observed…*

The same NLP techniques are applied to **generalize over semantic relations** that are not strictly equivalent. For instance sentences such as

> *The risk will be increased by comorbidities*

or

> *"Comorbidities should increase the risk.*

are not equivalent to our initial sentence, still, for the sake of generality, AsI-Health treat then in the same way (i.e. it extracts `comorbidity-increase-risk`). We are aware of the fact that in some situation this might be an oversimplification, but for the time being we prefer to favor high retrieval to high precision: indeed behind AsI-Health there is always an expert able to make the distinction between the different nuances. In the future we are likely to deal with these aspects in terms of features, such as modality, tens, etc.

Negation, on the contrary, is dealt this directly by changing the relation name. So

> *Comorbidities do not increase the risk.*

will become something as `comorbidity-increase_not-risk.`

Finally, we would like to point out the fact that arguments of relations are not always nominal in nature, even though this is the most frequent case. In complex sentences such as

We believe that comorbidities will significantly increase the death risk

we will have (at least) two relations, namely `comorbidity-increase-risk` and `we-believe-increase`. Of course we are aware of the fact that this is a poor treatment of intensional contexts, but for the time being this is what we can do to deal with such a complex topic. In the future we hope to be able to cope with the phenomenon either via reification or via featurization.

1.2 Relations with Prepositions

Of course semantic relations are expressed also by preposition, as in:

SARS-CoV-2 uses the same receptor for host cell entry.

In general the correct representation of this sentence would be something like "There is an action of SARS-CoV-2 using a receptor and this action is for entering host cell". Such a logical representation would however be too complex to be represented in the current data model and would require a sophisticated query language to be retrieved.

We decided therefore to make use of pseudo-relations and transform the above sentence in something like `SARS_CoV_2-use-receptor` and `SARS_CoV_2-use_for-entry`. The representation is redundant but it allows dealing with prepositional relations in exactly the same way as standard subject-object relations (notice that in the representation of triples the dash (-) represents the separator between subject, predicate and object, whereas the underscore (_) represent group of words grouped together in a single logical token.

1.3 Partial Relations

Partial relations are relations where either the subject or the object is absent. This might because one of them was really absent as in:

The contagion rate was reduced significantly.

In other cases the link is difficult to capture as in:

The contagion rate was reduced significantly but in order to be stopped by the current measures a different control system should be deployed…

Where it is clear the logical object of *stop* is *rate* but we can currently only extract something like `measure-stop`.

Finally we can have uncomplete relations due to parsing errors (missed identification of subject/object)

We decided to keep these binary relations as they might allow the identification of useful pieces of information, despite of the fact the information is not complete.

1.4 Nominals

In AsI-Health we will call "nominals" expressions occurring as subjects or objects. A crucial point in open information extraction is "to which degree of specificity should my nominals be represented?". Let's take a sentence such as:

Most patients had visited or worked in a seafood wholesale market in Wuhan.

Consider the relation whose predicate is *visiting*. Should the object be "*market*", "*wholesale market*" or "*seafood wholesale market*"? There is no clear-cut answer. Of course in logical terms we should have a set of atomic predicates representing something as "patients visited a market and the market is of type wholesale and the market is of seafood". As in the case of prepositional relations, however, this representation would be too complex for our model and difficult to query. We decided therefore to "explode" (or "unpack") the nominals and to keep them all. Some of the relations extracted from the above sentence would therefore be:

Subject	Predicate	Object
patients	visit	Market
patients	visit	seafood wholesale market
patients	visit	wholesale market
patients	visit	seafood market

With these triples the researcher is able to refine her search and ask, for instance "how many times papers report that future patients visited **a market**". However if she is rather interested in retrieving information about future patients visiting **seafood market**, our "unpacked" representation would equally provide answers.

Of course when "unpacking" is applied both to the subject and to the object we end up with a number of relations which is the Cartesian product of the unpacking results of the two. This is one of the reasons why in AsI-Health there are so many relations and why we had to rely on a rather powerful and scalable backend, i.e. ElasticSearch.