Fine Tuning Named Entity Extraction (DRAFT)

Paolo Curtoni
Innoradiant
Grenoble, France
paolo.curtoni@innoradiant.com

Luca Dini
Innoradiant
Grenoble, France
luca.dini@innoradiant.com

Abstract

In this report we provide a short workflow of the methodology applied in the context of Covid-19 MLIA @ Eval Initiative .

1 Introduction

In the context of the experiment we were interested in performing two tasks:

- Accurate, grammar-based identification of behaviors associated to Covid-19 pandemic.
- Massive identification of medical named entities.

In the following of this report we will describe the methodology applied to both tasks.

2 Grammar based behavior recognition.

For this task we configured a set of ad-hoc extractors which allowed the identification of different types of behavior. Each extractor is in turn composed of a set of rules ranging over triples in the RDF sense.

The triples serving as input to the extraction subsystem are produced by an application independent module available at Innoradiant called "Semarillion". Basically it takes as input the output of a dependency parser (in this case the Stanford ones) and applies a set of simplifications bridging the syntactic nature of the parsing (even though the Stanford parser already produce parse trees which are closer to semantic representations than traditional dependency trees) to a semantic representation more suitable for the application of future logic-oriented algorithm. Such simplifications include for instance, undoing of passivization, simplification of modal verbs chains (raising and control), featurization of prepositions, etc.

As each triple is represented by a subject a predicate and an object, rules are nothing more than constraints on lemmas of words appearing in these roles. Globally the grammar is composed of six different extractors and a total of 16 rules.

Rules where partially induced by representing the whole corpus as a graph where lemma identity corresponded to node identity and then running an algorithm (to be described) generalizing over handwritten triples. The generalization was performed by measuring cosine similarity among vectors of candidates and seed lemma. Vectors are obtained via a standard Word2Vector resource built on the specific Mlia corpus.

Unfortunately the approach suffered of many drawbacks:

- For many texts we could not obtain a reasonable dependency representation. Without such a representation the system is unable to produce any output.
- At configuration phase there was the assumption that the corpus contained "emotional behavior", i.e. subjective reactions to the different constraints, suggestions etc. Unfortunately these subjective attitudes are extremely rare, which makes the output far from satisfying on the recall side.
- We have the suspect that in the different steps some offset screwed up, thus generating inconsistent annotations.

3 Massive identification of medical named entities.

Rather than starting from scratch, we decided to adapt an existing medical named extraction system, namely Apache cTAKES (https://ctakes.apache.org/). We used it in its default configuration, without performing any retraining. It is worth here to say that in our configuration we activated the UMLS Terminology Services (https://uts.nlm.nih.gov/) in order to benefit of the most recent UML updates.

CTAKES was explicitly designed to analyze medical record clinical texts. As the Covid-19 corpus is not composed exclusively of such texts, the biggest disadvantage of cTAKES was an overgeneration of medical annotation in generic

context such as political news, standard reports etc. In order to contrast this tendency we built a classifier of MEDICAL vs GENERIC sentences and we decreased the degree of certainty emitted by cTAKES in GENERIC contexts.
The classifier was built on the basis of the corpus of the Multilingual Search track @MLIA, where a distinction between generic and medical texts is made. We randomly sampled English documents from “EU Press Corner”, “EUR-Lex” for the GENERIC class and from MEDISYS for the MEDICAL one. The model was obtained using standard BERT (bert-base-cased) with no hyper parameter tuning.

Beside the classifier, the other two building blocks used for fine-tuning the output are a terminology extraction system (TermSuite, Cram & Daille 2016) and a word2vec resource computed on the whole task1 corpus. The two modules are used both for enlarging the coverage and for limiting the overgeneration of cTAKES, that even on purely medical texts (or so classified) introduced many false positives.

The basic idea is that the terminology could be used on the one hand to validate the output of cTAKES, and on the other hand to inject terms that are medical in nature, but not necessarily in a technical sense (thus absent from cTAKES). For the first step we just filtered out all terms tagged as NE by cTAKES, but not flagged as terminological entries.

As for the injection of new entries (notably behaviors, which are not accounted at all in cTAKES) we adopted a seeded approach analogous to the one described in the previous section. For the behavior category we manually selected ten relevant terminology entries and used them as seed. These seeds were then matched against all other entries in the corpus-based terminology and the ones whose similarity with the seed was superior to a certain threshold where considered as **markers** for identifying behavior NE in the corpus. Currently markers are just regular expressions on syntactic tokens (POS and lemma) derived from above-threshold entries, with some generalization related to their syntactic nature (for instance a term of the form Noun-prep-Noun might be generalized as Noun-prep-(det|adj)*-Noun, if there is enough evidence in the corpus).

As for markers intended to increase the recall of the cTAKES filtered output, they were obtained with the same procedure, with the exception that rather than using a manually crafted seed, we selected as seed the first ten entries for each category where confidence of cTAKES was the highest.

4 Conclusions

Many improvements can be conceived to the method presented here, concerning both the computation of the vector associated to terms and the matching algorithm for multiword expressions. With the availability of a gold standard we will work in this direction and we will experiment different thresholds used in several modules of the system.

References

Damien Cram and Béatrice Daille. 2016. TermSuite: Terminology Extraction with Term Variant Detection. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics—System Demonstrations*, Berlin, Germany, 13–18.

Luca Dini, Paolo Curtoni and Elena Melnikova. 2018. Portability of Aspect Based Sentiment Analysis: Thirty Minutes for a Proof of Concept. Submitted to: *The 5th IEEE International Conference on Data Science and Advanced Analytics*. DSAA 2018, Turin.

Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*

Tomas Mikolov, Wen-tau Yih and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. *Proceedings of NAACL-HLT 2013*, 746–751.