Datasets

1. DISTANT-CTO is a weakly-labeled dataset of 'Intervention' and 'Comparator' entity annotated sentences. The dataset obtained using candidate generation approach described in "DISTANT-CTO: A Zero Cost, Distantly Supervised Approach to Improve Low-Resource Entity Extraction Using Clinical Trials Literature".
2. Physio test set is a dataset comprising 153 PICO annotated randomized controlled trial abstracts from Physiotherapy and Rehabilitation. This dataset was used as an additional benchmark to evaluate the generalization power of the weakly annotated dataset and NER model for this sub-domain.

Utility

The dataset could be used as an input for training 'Intervention' named-entity recognition (NER) models.

Availability

This directory includes **extraction1_pos_posnegtrail_conf09.txt** - This text data file contains all the weak annotations (source intervention terms mapped onto target sentences) from clinicaltrials.org (CTO) with a confidence score of 0.9 and above.

The directory also includes '**physio_sent_annot2POS_posnegtrail.txt**' – This data file contains manually annotated (Intervention entity) data from the physiotherapy and rehabilitation domain. It follows roughly similar structure as described in 'Description for long targets' section. ('Participant' and 'Outcome' annotations are removed from this file)

File Structure

NOTE: If you would like to fiddle with the example yourself, please open the **example.json** file in http://jsonviewer.stack.hu/.

The .txt data file consists of several lines, each line is stored in a JSON (short for JavaScript Object Notation) object representing one CTO record and the weak annotations obtained from this record. The topmost JSON object from each line consists of a 'string:value' pair containing the unique CTO ID of the CTO record (For example 'id:NCT04603443'). There are two nested JSON objects under the root json object with string '*extraction1*' and *'aggregate_annot*'. *'aggregate_annot*' contains all the annotations from '*extraction1*' justin aggregated form. As the project uses *'aggregate_annot*' for input, it's structure is described below (see Figure 1).

Under the '*aggregate_annot*' JSON object are the 'Intervention' entity-annotated targets *t*. The short targets (comprising only a single sentence) are arranged into an array while the long targets (comprising more than one sentence) are further arranged into a JSON object (see Figure 1).

Note: The submitted dataset file for the inspection contains only the *'aggregate_annot*' field due to the size limit for uploading supplementary material. When the data is released on GitHub, the entire file with '*extraction1*' fields will be made open access.

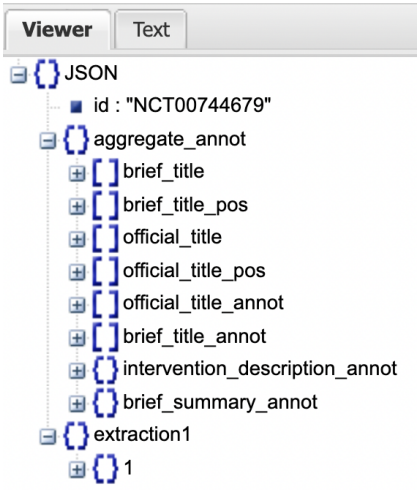

Figure 1: Example of a weakly-labeled CTO record stored as a JSON string in **extraction1_pos_posnegtrail_conf09.txt** file

Description for short targets

Each short target has its own 'targetname' which is a list of tokens from the tokenized target, 'targetname_annot' which is a list containing annotation for each individual token from the tokenized target, and 'targetname_pos' which is a list containing part-of-speech tags for each individual token from the tokenized target. An example is shown in Figure 2.

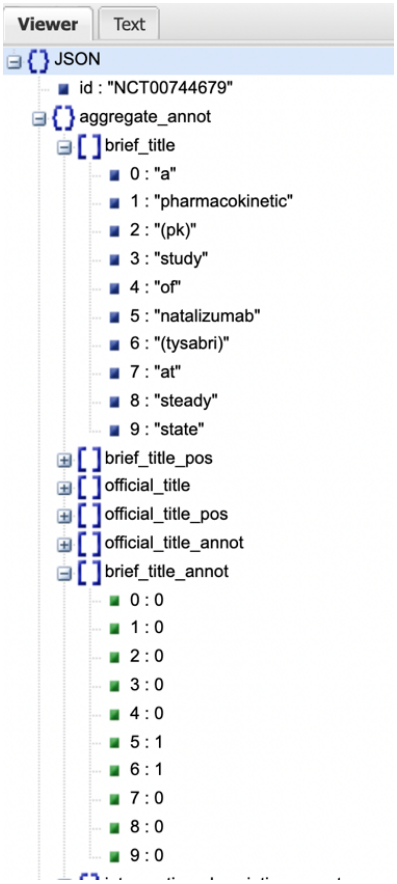

Figure 2: Example shows the structure of short targets tokens and their annotations.

Description for long targets

Annotation for each long target in a JSON object is flanked by the '_annot' keyword. Each long target which is a paragraph is tokenized into sentences and each sentence is stored as an array under the long target JSON object. Each sentence array is further divided into three lists. List 0 is a list of tokens from the tokenized sentence, list 1 is a list containing annotation for each individual token from the tokenized sentence and list 3 is a list containing part-of-speech tags for each individual token from the tokenized sentence. An example is shown in Figure 3.

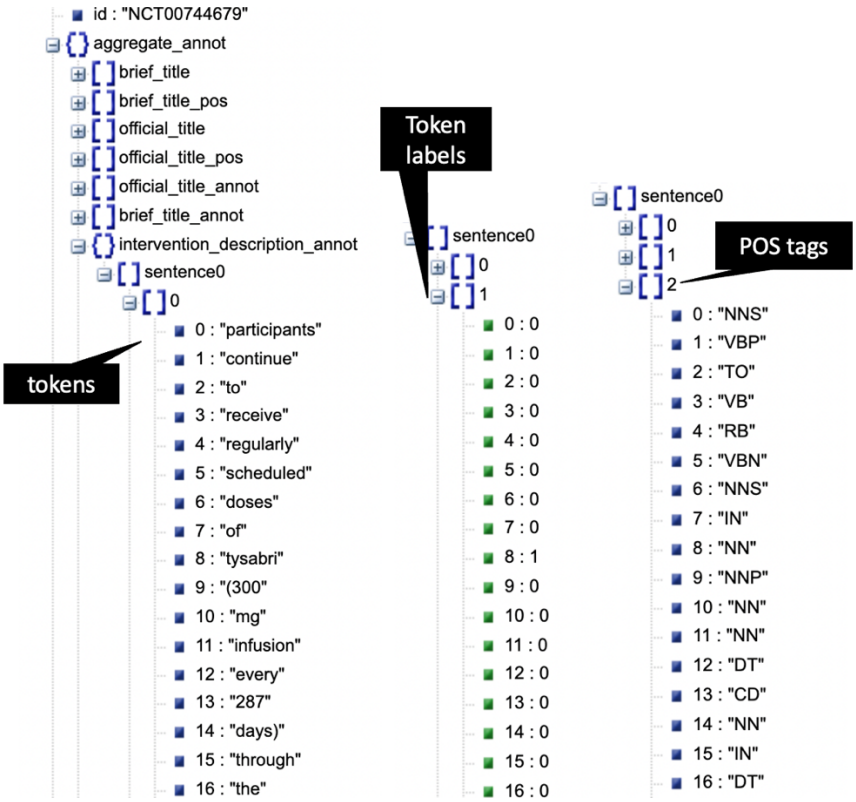

Figure 3: Example shows the structure of long targets tokens, their annotations and POS tags.

Ethical Statement

This paper studies clinical NER with a small strongly labeled and a large weakly labeled dataset. Our investigation neither introduces any social or ethical bias to the model nor amplifies any bias in the data. We do not foresee any direct social consequences or ethical issues.