Sub Classification NER Methodologies

2023-09-20

Guy Divita

# Sub Classification NER Methodologies and Efficacies

This section describes the NER algorithm and workflow methodologies and efficacies for the task of identifying Interpersonal Interactions and Relationships (IPIR) and Communication and Cognition (Comcog) subclasses, given sentences that have prior already classified as an IPIR and/or Comcog. All three IPIR and Comcog are sub trees of the International Classification of Functioning, Disability and Health (ICF) Activities and Participation Chapter. Relevant sections of those trees have been aggregated, binned and augmented to suit this SSA use-case.

Existing functionality from the Mental Functioning Ontology (MFO) Named Entity Recognition (NER) task has been extended and augmented to develop the capability to identify the IPIR and Comcog subclasses for this use-case.

The MFO NER includes an ontology that includes the components of the ICF Activities and Participation tree augmented with UMLS terminologies related to A&P and mental functioning along with related contextual concepts. The leaf nodes of the ontology overlap the subcategories pretty well. The existing NER functionality already marks phrases and terms with the ontology concepts.

This effort involves the following: *Terminology Curation*, *software development*, *manual annotation tasks*, *NER tuning*, followed by *evaluation*. Additional parts of that workflow, not yet done include a *feedback loop*, *retraining* and final *product deployment*.

## The Existing Technology Stack

The subclassification task is built from the MF Ontology NER, which is built from *java-nlp-Framework*[], functionality which evolved from v3NLP Framework, a suite of tools developed at the VA Salt Lake City to work with VA Clinical data. All of this is built on top of the Apache UIMA-Fit skin over the Apache UIMA NLP platform[]. Both UIMA and NLP-GATE architectures center around techniques that work with *stand-off annotations* within pipelines that pass containers holding the original text along with the stand-off annotations from one machine annotator to another. The *java-nlp-framework* functionality also integrated NLP-GATE functionality natively to read in and write out GATE formatted data.

An already existing Document Decomposition pipeline within the java-nlp-Framework is being employed to identify sentences, terms, tokens, section names for the text being processed. However, the sentence boundaries from the prior identified ipir\_yes and Comcog\_yes sentences are swapped in and used to keep the same segmentation across all of the inFACT processing.

The important component of this pipeline is the term identification within the text. Terms identified are based on dictionary/lexica lookup, where care has been taken to craft the content of the dictionaries, and where each of the dictionary entries contain term attributes including semantic categorization(s) for each term and the pedigree of where the term was sourced from. It is the semantic categorizations for each term identified within the text that is paramountly used down pipeline. While there are many ways to identify terms within the text, how and what is contained within the dictionaries is the salient part to the down pipeline machine annotators.

Additional machine annotators are employed to create new mental functioning classified annotations along with supporting contextual evidence annotations including adding Emotion evidences and Behavior evidences within each sentence. This process is over-generative and ad-hoc filtering has been added to the pipeline to stem some of the non-relevant annotations.

## Software Development

A new machine annotator (aka component of the pipeline) has been created for each of the 26 subcategories, where each of these pipeline components creates a subcategory evidence *annotation* for the span of the evidence found, and a new *sentence span mention annotation.* That is, at least two new machine annotations for each subcategory, one that spans the whole sentence, and one or more that spans each piece of evidence.

Once each of the sub-category machine annotators have been run, a summary machine annotator goes through each of the IPIR\_yes and Comcog\_yes sentences and affirmatively sets the subcategory attributes based on observing any of subcategory mentions within the sentence boundaries.

While most of these pipeline components operate independently of each other, the IPIR interaction and d779\_other subcategory machine annotators are ordered after the other IPIR machine subcategory annotators as these rely on the presence or absence of other IPIR evidence. For the Comcog subcategories, the SSA attribute classes including Applied Memory, adaptation, persistence and pacing are run after the other Comcog pipeline components, as these require the presence of other Comcog evidences as prerequisites to attach those attributes to.

There is a check to throw out mentions that are not IPIR, Comcog, or Mental Functioning related due to some known counter-factual context. For example, if the sentence has the word invoice, or billing in it, this is not an MFO mention. Within the IPIR terminologies, there are 162 terms marked as not IPIR related. More examples include balance support, math support, and with respect to. There are an additional 1389 terms marked as NotMFO which includes case number, department of social services, social security act along with terms like swelling and please. These non-relevant evidence terms were garnered and added from frequency distributions of terms appearing in false positive mentions and never appearing in true positive or false negative mentions.

A final heuristic universally has been added that sets the *Other* category to true when there is no underlying evidence found. It was deemed safe to do so, in part, because the task at hand already knows that the sentence being processed has already been identified as an asserted case from an up pipeline inFACT process.

## Terminology Curation

The *java-nlp-framework* includes by default, the 2022 UMLS SPECIALIST Lexicon that covers general English, and curated focused terminologies to cover terms within clinical section names, labs, units of measure, person evidence etc.

For this sub classification work, additional lexica were curated. Terms thought to be the most general IPIR and Comcog classes from the MF Ontology were gathered. Each of these were used as seed terms to traverse through the 2022 UMLS to gather all descendent terms to be incorporated into the terminology used for lookup. IPIR and Communication and Cognition activities are poorly covered in the UMLS, and this process was disappointing, only picking up ICF Activities and Participation subclasses, for the most part.

It was recognized that communication and cognition activities evidence are likely to be verbs and adverbs. An effort was made to classify the University of Colorado’s VerbNet classes into IPIR and Comcog categories manually. That effort was expanded to have each of the IPIR and Comcog categories (about 162 classes) to further be subclassified into the 20 subclasses manually. This was done by 4 domain expert annotators and one software engineer (me). This was a daunting task involving classifying over 5000 lemmas resulting from the 162 classes. There was reason for classifying at the lemma level rather than class level. The classes were too inclusive, with most classes involving lemmas (in VerbNet terminology: members) from the classes that didn’t really fit the semantic relatedness criteria we were classifying at. We had time constraints so the effort was a quick and rough effort with the understanding that we’d go back and do better curation when time allowed. For the most part, the VerbNet Comcog tagged lemmas is the terminology employed.

Additional terminologies were sought for IPIR classes to garner role and authority evidence, taken from the Bureau of Labor Statistics public sources, evidence around strangers, intimate relationships. The sources to family relationships came directly from terms marked as family history from the UMLS. Other lexica around informal relationships, interactions, and such were garnered through introspection, and online thesaurus to get related terms.

## Manual Annotation

This task was fortunate to have access to manual annotations for most of the subclasses. The subclassification annotation task is described elsewhere.

## NER Tuning

Once the terminology was roughly in place, the NER was coded and the training set were then used to tune each subcategory, addressing the more prevailing issue: more false positives than false negatives or visa-versa. For false negatives, term frequencies from the missed mentions were created, the top (manually recognized) salient terms from the list were considered to be added to the lexicon, where each term added was evaluated to see if the added term added significant false positives or not. For false positives, terms from the top frequency false positive evidence were removed when their removal did not incur too many false negatives as a consequence. Removed is an overstatement: for the term in consideration, the subcategorization category was commented out, so the term, while still known, would not be used to classify that subcategory.

# Evaluation/Efficacy

The efficacy for Comcog subcategories are in table 1. The efficacy for IPIR subcategories are in table 2.

## Failure Analysis

Noted: a small but obvious portion of the false negatives were from *Slot:Values* where the answer part to the slot or question was in a second segment. Answers, lacking the full context, were universally not marked. This was an engineering trade-off, to write-off those failures knowing they could be addressed, because the gain was small. I never got the cycles to write a wrapper to check whether a \_yes sentence is within a *Slot:Value*.

Noted: only a small number of false negatives were the result of out-of-terminology evidences. Out-of-terminology kinds of failures are easy to remedy when seen. It’s a question of training exposure to pick these up.

Noted: A significant number of false negatives were the result of non-obvious inferences. When there is time, this class of failures should be followed up in a feed-back loop with the annotators to verify these were truly positive cases, and to gather insight how to identify the context around these cases. For these classes of failure, this simplistic term-based lookup technique may not be appropriate.

## Discussion

The training annotations were skewed to majority class with little to no coverage for attributes that would be most helpful for the SSA use case. In particular, *adaptation*, *persistence*, *pacing*, and *d110\_d129\_Learning and Applying Knowledge*. As such, the efficacy of these attributes are weak because the cell sizes are too low. We did gather terminology for these classes, ergo in theory, a follow-up would be to augment the existing samples with pages pulled using terms we know to be from these classes.

Only a limited amount of time was taken to tune these classifiers. Efficacy targets are within achieving given more time and exposure to more targeted manual annotations.

[TBD – pending article submission ]