General notes for interpreting the Java Indicator Rules file for the SemRep Program

SemRep (https://semrep.nlm.nih.gov/) is a rule-based natural language processing program. It uses structured domain knowledge in the UMLS (https://www.nlm.nih.gov/research/umls/index.html) to represent textual content as three-part *semantic predications*. These are propositions that SemRep extracts from assertions in sentences of biomedical text, preferably in Medline format. *Semantic predications* consist of two UMLS Metathesaurus (https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html) concepts as arguments and a UMLS Semantic Network (SN) (https://semanticnetwork.nlm.nih.gov/) predicate that names the nature of their relationship. For example, from (1), SemRep extracts the predication in (2):

(1) *Respiratory infections are often accompanied by mild hepatitis*. [PMID=31505263]
(2) Hepatitis COEXISTS_WITH Respiratory Tract Infections

Indicator rules

While the UMLS provides the predication arguments and the linking predicates, indicator rules map syntactic elements in the text, such as verbs and nominalizations, to predicates in the SN (e.g., TREATS, PREVENTS, AFFECTS, and so on). For the predication in (2), for example, the indicator rule needed is (3).

(3) *accompanied_by* (verb) → COEXISTS_WITH

Why indicator rules are needed

Since there are many different ways to express a given thought in natural language, the indicator rules capture the fact that the associations or interactions between entities (arguments) can be expressed in any number of ways but the meaning may be the same. For instance, some indicators that map to the SN predicate TREATS are *treats, management, treatment, control,* and *ameliorate*, used in the following:

Aspirin treats headaches.
Statins treatment/management for hypertension.
Metformin is recommended for diabetes control.
Non-steroidal anti-inflammatory drugs (NSAIDs) may alleviate the pain.

Conversely, a given syntactic expression may indicate different meanings, representing different semantic predicates. The verb *abate*, for example, may represent either AFFECTS or DISRUPTS depending on the semantic nature of the entities with which it is used. Thus, *abate* is an indicator for both. Entities are mapped to their corresponding semantic concepts through the UMLS Metathesaurus. Indicator rules apply to entire semantic categories, not individual entities, generalizing across classes of nouns.

How to read the Java Semrules file

This file contains the SemRep indicators for the SN predicates that SemRep uses. At this time (version v1.7), SemRep is using an earlier Prolog file. The new JAVA file was created for use with SemRep v1.8 in a JAVA implementation. Thus, some rules have not been implemented yet but are included in the file for future implementation after SemRep Java is complete.
There are different types of indicator rules, from simple to multi-phrase, and the format varies for each. However, as illustrated in Figure 1, all rules contain:

a) The **indicator:** a word or words that may appear in texts.

b) The indicator's **part-of-speech** (POS), written as a two-or three-letter abbreviation (VB, NN, JJ) Penn TagSet (https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html)

c) A **cue** that may introduce the object argument, usually a preposition, but it can be more complex. In the expression ***synergy between A and B,*** the indicator is the noun *synergy*; its two arguments are to its right, preceded by preposition *between,* conjoined by *and*. This text fragment would generate A INTERACTS_WITH B, as *synergy* maps to predicate INTERACTS_WITH, with **cues** '*between-and*'.

d) The ***category*** or SN predicate to which the indicator maps. The *category* is always the non-passive, non-negative form. For example, the passive indicator *treated_by* maps to TREATS.

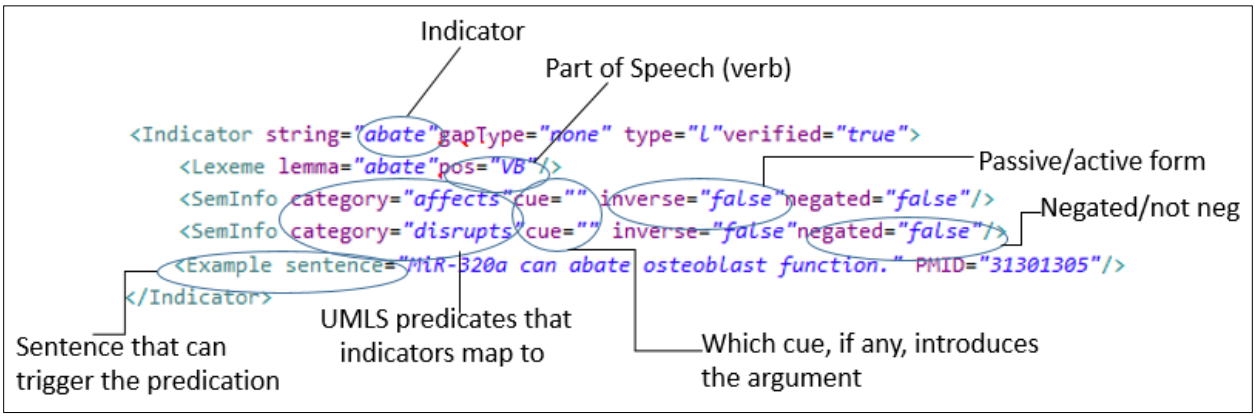
e) The **passive form** is indicated as 'inverse=”true”'.

f) **Negation** is indicated as 'negated=”true”'.

g) For **multiword**, **multi-phrase** indicators, the head for argument identification is placed first.

h) When possible, **examples** are provided to test the indicator rules (**Example** element). If there are no examples, a comment to that effect was added in the file.

Figure 1. Screenshot illustrating indicator *abate*, which maps to SN predicates AFFECTS and DISRUPTS.



Sources used for the example sentences

Most example sentences come from PubMed, and the PMID number is provided at the end of the sentence line. Other sentences come from general internal files or were conceived with the allowed arguments in mind, to favor the generation of the desired predication(s). Many sentences were simplified or reworded to facilitate the desired result.

Additional role for the example sentences

The collection of sentences can serve as a Gold Standard for what SemRep is expected to generate after JAVA implementation is complete. It is valuable for version control of different SemRep versions and new releases of the UMLS Metathesaurus, which SemRep crucially depends on. Changes in newer versions and/or releases may affect SemRep's output, which would be reflected in the Gold Standard results.

December 2019