A Guide to Understanding SemRep Full-Fielded Output

Halil Kilicoglu

1 Introduction

The output of SemRep full-fielded processing falls into four categories:

- `text`
- `entity`
- `relation`
- `coreference` (Only with anaphora_resolution option, -A)

All fields are separated by "|"; certain fields (preceded by "*" below) can be empty, although in non-production output, they may be represented by non-empty placeholders, as described below.

2 Fields Common to All Output

All output will have the same first five fields:

1. `SE`: designates that the output is from SemRep.
2. PMID
3. * Subsection: If the utterance begins with one of a specified set of strings of uppercase letters followed by a colon (see Appendix A for a complete listing of these strings) this field will contain that string; otherwise it is blank.
4. `ti` if the utterance is from the title of the citation; `ab` if the utterance is from the abstract of the citation.
5. Sentence ID: an integer indicating the utterance's position within the title/abstract.

3 Sixth Field

The sixth field indicates the output type, and will be one of atoms itemized in Section 1: `text`, `entity`, `relation`, or `coreference`.

4 Remaining Fields for text Output

A typical line of text output looks like this:

```
SE|15311027|RESULTS|ab|12|text|256|296|No major complications
were experienced.
```

Text output contains 9 fields. The first 6 fields were described in Section 2; the rest of the fields are as follows:

7. First character position (in document) of the utterance.
8. End position (in document) of the utterance.
9. The ASCII text of the utterance.

5 Remaining Fields for entity Output

A typical line of entity output looks like this:

```
SE|17208639||ti|1|entity|C0027893|neuropeptide Y|aapp,nsba|4852|
NPY|neuropeptide y|||0|1000|39|59
```

Entity output contains 16 fields. The first 6 fields were described in Section 2; the remaining 11 fields are the following:

7. * CUI of the entity[1] (`C0027893`)
8. * Preferred name of the entity[2] (`neuropeptide Y`)
9. Semantic Type(s) of the entity[3] (`aapp,nsba – Amino Acid, Peptide, or Protein and Neuroreactive Substance or Biogenic Amine` in the example above)
10. * Normalized gene ID(s) from EntrezGene; may contain multiple IDs delimited by comma or may be empty (`4852`)
11. * Normalized gene name(s) from EntrezGene; may contain multiple names delimited by comma or may be empty (`NPY`)
12. Text in the utterance that maps to the entity (`neuropeptide y`)
13. * Change term; empty in SemRep output (`<CHANGE>` may appear as placeholder)
14. * Degree term; empty in SemRep output (`<DEGREE>` may appear as placeholder)
15. * Negation term: 1 if the entity is negated, 0 if it's not. (`0`)
16. Confidence score (integer between 0 and 1000; rarely below about 250) (`1000`)
17. First character position (in document) of text denoting entity (`39`)[4]
18. End position (in document) of text denoting entity (`59`)

6 Remaining Fields for relation Output in SemRep

A typical line of SemRep relation output looks like this (the line is broken for readability; in the actual output, all text will appear on one line):

```
SE|00000000||tx|1|relation|3|1|C0027893|neuropeptide Y|
aapp,gngm,nsba|aapp|4852|NPY|neuropeptide y|||0|1000|39|59|
VERB|INHIBITS||70|79|4|2|C0021753|Interleukin-1 beta|
aapp,gngm,imft|gngm|3553|IL1B|interleukin-1beta|||0|1000|129|136
```

SemRep Relation output contains 41 fields. The first 6 fields were described in Section 2; the remaining 39 fields are the following:

7. SubjectMaxDist: The number of potential arguments (i.e., NPs) from the indicator in the direction of the subject (`3`)
8. SubjectDist: The number of potential arguments separating the subject from the indicator (`1`)
9. * CUI of the subject concept (`C0027893`)
10. * Preferred name of the subject concept (`neuropeptide Y`)

---

[1] Entities extracted only from EntrezGene will not have CUIs.
[2] Entities extracted only from EntrezGene will not have MetaConcs.
[3] Entities extracted only from EntrezGene will have 'gngm' (Gene or Genome) as their Semantic Type.
[4] All character offsets in SemRep full-fielded output are 0-based.

11. Semantic Type(s) of the subject concept[5] (`aapp, gngm, nsba` in the example above, `gngm` is an artificial semantic type)
12. Subject Semantic Type used for the relation (`aapp`)
13. * Normalized gene ID(s) of the subject from EntrezGene; may contain multiple IDs delimited by comma or may be empty (`4852`)
14. * Normalized gene name(s) of the subject from EntrezGene; may contain multiple names delimited by comma or may be empty (`NPY`)
15. Text that maps to the subject (`neuropeptide y`)
16. * Change term (`<CHANGE>` may appear as placeholder)
17. * Degree term (`<DEGREE>` may appear as placeholder)
18. * Negation term: 1 if the subject is negated, 0 if it's not. (`0`)
19. Confidence score (`1000`)
20. First character position (in document) of text denoting subject entity (`39`)
21. Last character position (in document) of text denoting subject entity (`59`)
22. Indicator Type[6] (`VERB`)
23. Predicate (`INHIBITS`)
24. `negation` if the relation (the immediately preceding field) is negative; empty otherwise
25. First character position (in utterance) of text denoting relation (`70`)
26. End position (in utterance) of text denoting relation (`79`)
27. ObjectMaxDist: The number of potential arguments (i.e., NPs) from the indicator in the direction of the object (`4`)
28. ObjectDist: The number of potential arguments separating the object from the indicator (`2`)
29. * CUI of the object concept (`C0021753`)
30. Preferred name of the object concept (`Interleukin-1 beta`)
31. Semantic Type(s) of the object concept (`gngm,aapp,imft` in the example above, `gngm` is an artificial semantic type)
32. Object Semantic Type used for the relation (`gngm`)
33. * Normalized gene ID(s) of the object from EntrezGene; may contain multiple IDs delimited by comma or may be empty (`3553`)
34. * Normalized gene name(s) of the object from EntrezGene; may contain multiple names delimited by comma or may be empty (`IL1B`)
35. Text that maps to the object (`interleukin-1beta`)
36. * Change term (`<CHANGE>` may appear as placeholder)
37. * Degree term (`<DEGREE>` may appear as placeholder)
38. * Negation term: 1 if the object is negated, 0 if it's not. (`0`)
39. Confidence score (`1000`)
40. First character position (in document) of text denoting subject entity (`129`)
41. End position (in document) of text denoting subject entity (`136`)

---

[5] Some of these semantic types may be artificial. For instance, SemRep adds the semantic type 'gngm' (Gene or Genome) if the original semantic type is 'aapp' (Amino Acid, Peptide, or Protein) and vice versa.

[6] Possible values: `PREP` (preposition), `MOD/HEAD` (intra-NP relation), `VERB` (verb), `NOM` (nominalization), `SPEC` (hypernymy), `INFER` (inference)

7 Remaining Fields for coreference Output in SemRep

A typical line of SemRep `coreference` output looks like the following:

```
SE|15996060|OBJECTIVE|ab|3|coreference|C0019932|Hormones|horm|||
these hormones||||1000|694|708|COREF|C0014939|Estrogens|
horm,phsu,strd|||estrogens||||1000|582|591
```

Fields 7-18 correspond to `anaphor` element of the `coreference` relation and the fields 20-31 correspond to the `antecedent` element. Both `anaphor` and `antecedent` are entity objects, and their individual fields correspond to the elements of the entity object, as described in Section 5.

Appendix A: Subsection terms

The third field of the full-fielded output lines may refer to the subsection of the text. We compiled a list of subsections from Medline abstracts. Currently, this list contains approximately 8,000 section names. Here is a partial list

ANIMALS
AVAILABILITY
BACKGROUND
BACKGROUND AND AIMS
BACKGROUND AND OBJECTIVE
BACKGROUND AND OBJECTIVES
BACKGROUND AND PURPOSE
CASE REPORT
CLINICAL IMPLICATIONS
CLINICAL RELEVANCE
CONCLUSION
CONCLUSIONS
CONCLUSIONS AND CLINICAL RELEVANCE
CONTEXT
DATA COLLECTION AND ANALYSIS
DATA SOURCES
DATA SYNTHESIS
DESIGN
DESIGN AND METHODS
DESIGN AND SETTING
DEVELOPMENT
DISCUSSION
EXPERIMENTAL DESIGN
FINDINGS
HYPOTHESIS
IMPLICATIONS

IMPLICATIONS FOR NURSING PRACTICE
INTERPRETATION
INTERVENTION
INTERVENTIONS
INTRODUCTION
LIMITATIONS
MAIN OUTCOME MEASURE
MAIN OUTCOME MEASURES
MAIN RESULTS
MATERIAL AND METHOD
MATERIAL AND METHODS
MATERIALS AND METHODS
MEASUREMENTS
MEASUREMENTS AND MAIN RESULTS
MEASUREMENTS AND RESULTS
MEASURES
METHOD
METHOD OF STUDY
METHODOLOGY
METHODS
METHODS AND MATERIALS
METHODS AND RESULTS
MOTIVATION
OBJECT
OBJECTIVE
OBJECTIVES
OUTCOME MEASURES
PARTICIPANTS
PATIENTS
PATIENTS AND METHOD
PATIENTS AND METHODS
POPULATION
PROBLEM
PROCEDURE
PURPOSE
PURPOSE OF REVIEW
PURPOSE OF THE STUDY
RATIONALE
RATIONALE AND OBJECTIVES
RECENT FINDINGS
RELEVANCE
RESEARCH DESIGN AND METHODS
RESEARCH METHODS AND PROCEDURES
RESULT
RESULTS
RESULTS AND CONCLUSIONS

SAMPLE
SEARCH STRATEGY
SELECTION CRITERIA
SETTING
SIGNIFICANCE
SIGNIFICANCE AND IMPACT OF THE STUDY
STATEMENT OF PROBLEM
STUDY DESIGN
STUDY DESIGN AND METHODS
STUDY OBJECTIVE
STUDY OBJECTIVES
STUDY SELECTION
SUBJECTS
SUBJECTS AND METHODS
SUMMARY
SUMMARY BACKGROUND DATA
SUMMARY OF BACKGROUND DATA