Linking UNASSIGNED and Unmapped Structured Abstract Labels to the Five Canonical NLM Categories

April 6, 2022

François-Michel Lang

1 Background

The 2022 MEDLINE Baseline contains 4,589,456 MEDLINE citations with Structured Abstracts (SAs)[1] (13.74% of the 33,405,863 citations in the Baseline) which in turn contain 19,430,720 instances of labels. 15,677,756 (80.69%) of these label instances (2,974 distinct, case-normalized labels) in 3,623,649 citations are well constructed in the sense that they are linked to one of the five canonical NlmCategories (BACKGROUND, OBJECTIVE, METHODS, RESULTS, and CONCLUSIONS). A few examples of such well-constructed labels are

```
<AbstractText Label="INTRODUCTION AND MATERIAL" NlmCategory="BACKGROUND">
<AbstractText Label="PURPOSE OF INVESTIGATION" NlmCategory="OBJECTIVE">
<AbstractText Label="DESIGN SETTING AND PARTICIPANTS" NlmCategory="METHODS">
<AbstractText Label="MEASUREMENTS AND MAIN FINDINGS" NlmCategory="RESULTS">
<AbstractText Label="CLINICAL SIGNIFICANCE" NlmCategory="CONCLUSIONS">
```

2 Objective

The goal of this project is to assign one of the five canonical NLM categories to structured abstract labels that are not well constructed because they have no such NlmCategory. These remaining 3,752,964 labels (= 19,430,720 − 15,677,756) fall into three categories, presented next.

2.1 Labels with NlmCategory of "UNASSIGNED"

First, 276,986 (1.42%) label instances (13,101 distinct after case normalization) in 100,035 citations have an NlmCategory of `UNASSIGNED`, e.g.,

```
<AbstractText Label="INTERVENTIONS IN BARS" NlmCategory="UNASSIGNED">
<AbstractText Label="AN OBSERVATIONAL STUDY" NlmCategory="UNASSIGNED">
<AbstractText Label="THE MODIFICATION OF RISK" NlmCategory="UNASSIGNED">
<AbstractText Label="IMPLICATIONS FOR HEALTH POLICY AND NURSING" NlmCategory="UNASSIGNED">
<AbstractText Label="MORE LESSONS FROM DEVELOPED COUNTRIES FOR IMCI" NlmCategory="UNASSIGNED">
```

[1] For this project, we ignore `<OtherAbstract>`s.

2.2 Unmapped Labels (no NlmCategory)

Furthermore, 3,391,616 (17.45%) of all label instances (6,927 distinct after case normalization) in 789,638 citations are unmapped because they have no NlmCategory at all, e.g.,

```
<AbstractText Label="THE ROLE OF HERBS AND SPICES IN HEALTH">
<AbstractText Label="DATA SOURCES, EXTRACTION, AND SYNTHESIS">
<AbstractText Label="PARTICIPANTS, INTERVENTION, AND MEASURES">
<AbstractText Label="PARTICIPANTS/MATERIALS, SETTING, METHODS">
<AbstractText Label="CONCLUSIONS AND IMPLICATIONS FOR PRACTICE">
```

Note that 2,477 distinct case-normalized labels appear in both UNASSIGNED and unmapped categories, e.g.,

```
<AbstractText Label="ACCESSION NUMBERS" NlmCategory="UNASSIGNED">
<AbstractText Label="Accession numbers">

<AbstractText Label="ACQUISITION OF EVIDENCE" NlmCategory="UNASSIGNED">
<AbstractText Label="ACQUISITION OF EVIDENCE">
```

In all, we dealt with 3,668,602 instances of 17,151 distinct case-normalized labels that appear in either an UNASSIGNED or unmapped XML element.

2.3 UNLABELLED Labels

Finally, we note for completeness that 84,362 labels (0.43%) in 84,362 citations have "UNLABELLED" as the label, e.g.,

```
<AbstractText Label="UNLABELLED">
```

but for this project, we ignore UNLABELLED labels.

3 Summary of Label and Citation Counts

Label Type	Instances	Distinct	Citations
Well formed	15,677,756	2,974	3,623,649
UNASSIGNED	276,986	13,101	100,035
unmapped	3,391,616	6,927	789,638
UNLABELLED	84,362	1	84,362
TOTAL with overlap	19,430,720	20,029	4,597,684
TOTAL w/o overlap	19,430,720	20,029	4,589,456

Note that the sum of the citation counts for UNASSIGNED, unmapped, and UNLABELLED citations (4,597,684) exceeds the number of structured abstract citations (4,589,456) noted above in section 1. This is due to

- 4,836 citations containing both an UNASSIGNED and UNLABELLED label, and
- 3,392 citations containing both an UNLABELLED and unmapped label

4 Methods

This document proposes a method of automatically linking UNASSIGNED (section 2.1) and unmapped (section 2.2) labels to one of the five canonical NlmCategories (section 1).

We present now three methods for linking most of the UNASSIGNED and unmapped labels presented in section 2, which, as noted above, include 3,668,602 instances of 17,151 distinct labels.

4.1 Automatic Linking

In 2015, NLM subject-matter experts produced a linking of 3,032 distinct structured abstract labels to one of the five canonical NlmCategories. We will refer to these linkings as the *2015 linkings*; they are available at

```
https://lhncbc.nlm.nih.gov/ii/areas/structured-abstracts/
        downloads/Structured-Abstracts-Labels-102615.txt
```

and look like

```
ETHICS AND DISSEMINATION|BACKGROUND|N|20131106
AIM OF THE STUDY|OBJECTIVE|N|20100629
MATERIAL AND METHODS|METHODS|N|20100629
ACHIEVEMENTS|RESULTS|Y|20151026
KEY MESSAGE|CONCLUSIONS|Y|20151026
```

These 2015 linkings allowed us to map to one of the five canonical NlmCategories both UNASSIGNED and unmapped labels such as

```
<AbstractText Label="ETHICS AND DISSEMINATION">
<AbstractText Label="AIM OF THE STUDY">
<AbstractText Label="MATERIALS AND METHODS" NlmCategory="UNASSIGNED">
<AbstractText Label="ACHIEVEMENTS">
<AbstractText Label="KEY MESSAGE" NlmCategory="UNASSIGNED">
```

Linking UNASSIGNED and unmapped label instances that appear in the 2015 list is automatic, and resulted in mapping 3,520,274 instances (95.96%) of 2,112 (12.45%) distinct labels.

4.2 Algorithmic Linking by Minimum Edit Distance

The first algorithmic-linking technique computes, for each label not automatically mapped, the minimum edit distance[2] to all label instances that had been automatically linked. Inspection of results revealed that minimum edit distances of

[2] See e.g., https://en.wikipedia.org/wiki/Edit_distance, https://observablehq.com/@stwind/minimum-edit-distance, and

- =1 were universally true positives; see items 1–3 immediately above;
- =2 required human review, but were mostly true positives 329 of 401 (82.04%);
- =3 also required human review, but nonetheless contributed many true positives 176 of 524 (33.59%).
- >3 were unreliable and ignored—essentially noise

This strategy was especially fruitful for linking

1. misspellings, e.g., `abstarct`, `backgroud`, `conslusions`, `ntroduction`, `objetive`, `pupose`;
2. non-English (principally Spanish and Portuguese) terms, e.g., `caso clinico`, `intervenciones`, `introduccion`, `resultados`, `conclusao`, `introducao`, etc., with close English cognates;
3. plural forms not in the 2015 list, but whose lexicographically similar singular form *does* appear in the 2015 list, e.g., `concepts`, `contributions`, `diagnoses`, `responses`.

Minimum-edit-distance linking (distance < 4) added 39,408 instances of 1,199 labels, bringing the totals after automatic and minimum-edit-distance linking to 3,559,682 (97.03%) instances of 3,313 (18.36%) labels.

4.3 Algorithmic Linking by Score

The final step for linking UNASSIGNED and unmapped labels is based on a scoring algorithm that relies in part on the contents of a January 23, 2014 NLM e-mail exchange (included for reference as an appendix) between NLM subject-matter experts which established a priority ranking of the five canonical NLM categories (section 1). We assigned numerical ranks to the five categories, from highest to lowest priority:

OBJECTIVE	5
CONCLUSIONS	4
RESULTS	3
METHODS	2
BACKGROUND	1

We take as an example the label `STATEMENT OF SIGNIFICANCE`, which appears 1,787 times in the 2021 Baseline.

4.3.1 Scoring Each Word in Label

We begin by examining the occurrences of each word in the label (excluding PubMed stopwords)[3]) as a token in the 2015 linkings. For example, `STATEMENT` appears 15 times:

`https://web.stanford.edu/class/cs124/lec/med.pdf.`

[3] `https://pubmed.ncbi.nlm.nih.gov/help/#help-stopwords`

```
CONCLUDING STATEMENT|CONCLUSIONS
CONFLICT-OF-INTERESTSTATEMENT|BACKGROUND
CONSENSUS STATEMENT|METHODS
IMPLICATION STATEMENT|CONCLUSIONS
IMPLICATIONS STATEMENT|CONCLUSIONS
PROBLEM STATEMENT|OBJECTIVE
PROBLEM STATEMENT AND BACKGROUND|OBJECTIVE
PROBLEM STATEMENT AND PURPOSE|OBJECTIVE
STATEMENT OF CONCLUSIONS|CONCLUSIONS
STATEMENT OF PROBLEM|BACKGROUND
STATEMENT OF PROBLEM AND RATIONALE|BACKGROUND
STATEMENT OF PROBLEMS|BACKGROUND
STATEMENT OF PURPOSE|OBJECTIVE
STATEMENT OF THE PROBLEM|BACKGROUND
SUMMARY STATEMENT|CONCLUSIONS
```

The category counts for those 15 lines are

OBJECTIVE	4
CONCLUSIONS	5
RESULTS	0
METHODS	1
BACKGROUND	5

We then perform the same calculation on `SIGNIFICANCE`, which appears 36 times in the 2015 list; the category counts for these 36 occurrences are:

OBJECTIVE	0
CONCLUSIONS	32
RESULTS	1
METHODS	0
BACKGROUND	3

4.3.2 Combining Word Scores

We then sum the category counts for each word:

OBJECTIVE	4 + 0	4
CONCLUSIONS	5 + 32	37
RESULTS	0 + 1	1
METHODS	1 + 0	1
BACKGROUND	5 + 3	8

and finally multiply each category's sum by its priority given above:

OBJECTIVE	4 * 5	20
CONCLUSIONS	37 * 4	148
RESULTS	1 * 3	3
METHODS	1 * 2	2
BACKGROUND	8 * 1	8

According to this analysis, CONCLUSIONS has the highest score (148) of the five canonical Nlm-Categories and is therefore the winning category. In case of ties (e.g., if OBJECTIVE and CONCLUSIONS had both scored 148), the winner would be the category with the higher priority.

Note that this algorithm does not identify a closest already-mapped label, but only the most likely canonical NLM category.

We then manually reviewed the 1,208 remaining labels whose

- score was at least 100, and
- frequency of occurrence was at least 2

and in certain cases, modified the canonical NLM category assigned by the scoring algorithm.

Linking by score added 43,335 instances of 1,002 labels, bringing the totals after automatic and minimum-edit-distance linking to 3,603,017 (98.21%) instances of 4,313 (25.43%) labels. After the three linking steps, we were left with only 65,585 (1.79%) unlinked instances of 12,838 (74.85%) distinct labels.

Remaining unlinked labels fall into several categories:

- Possible errors or highly technical or specific terms, e.g., “, ffw”, “pbh-lci”, “bw, adg, f”;
- Foreign terms, e.g., ergebnisse, einleitung, fortolkning with no close English cognate;
- Perfectly normal labels that unfortunately have no similarity to any previously linked label, e.g., drugs and falls, laser parameter, pleural lesions, lymphatic tissue, cellular viability, etc.; and
- Exceptionally long labels, e.g.,
  - gait mainly depends on the relationship between posture balance and movement,
  - linear versus sigmoid relationship between blood pressure fall and drug concentration,
  - polymerization shrinkage stress and stress reduction possibilities.

The table on the next page summarizes all the above results.

5 Conclusion

Notice that in the final chart below (Linked after AUTO, SCORE, and DISTANCE), the percentage of distinct labels linked (25.15%) is far lower than the percentage of label instances linked (98.21%).

This difference suggests a long tail of low-frequency labels that could not be linked to one of the five canonical NlmCategories. Indeed, labels that were successfully linked have an average frequency of occurrence of 835.39, whereas labels that could not be linked have an average frequency of 5.11.

Automatic and algorithmic linkings have served to link a vast majority of UNASSIGNED and unmapped labels. We hope that these results might lead to the linking of most structured abstract labels in the Baseline.

	distinct labels		label instances	
	count	%age	count	%age
Orig Unlinked	17,151	100.00%	3,668,602	100.00%

	Linked after AUTO			
	distinct labels		label instances	
	count	%age	count	%age
Linked	2,112	12.31%	3,520,274	95.96%

	Linked after AUTO and DISTANCE			
	distinct labels		label instances	
	count	%age	count	%age
Linked	3,113	18.15%	3,559,682	97.03%

	Linked after AUTO, DISTANCE and SCORE					
	distinct labels		label instances			
	count	%age	count	%age		
Linked	4,313	25.15%	3,603,017	98.21%	835.39	Avg frequency

Final Unlinked	12,838	74.85%	65,585	1.79%	5.11	Avg frequency

Appendix

In 2014, NLM subject-matter experts ranked the 5 canonical metadata NlmCategories (BACKGROUND, OBJECTIVE, METHODS, RESULTS, CONCLUSIONS) by order of importance for mapping structured abstract labels and information-retrieval experimentation. The rankings are as follows:

1. OBJECTIVE
2. CONCLUSIONS
3. RESULTS
4. METHODS
5. BACKGROUND

OBJECTIVE is ranked first because of the overall contextual importance of a study's purpose. CONCLUSIONS is ranked ahead of RESULTS because in the majority of cases, CONCLUSIONS does limit itself to the major results; some branch off into the further-research arena, but not enough to skew the main thrust of an article. Moreover, CONCLUSIONS is supported by results/discussion in the article. METHODS is ranked fourth because of the tendency to find non-semantic-type concepts indicative of patient population characteristics, experimental animals, and publication types. BACKGROUND is last because it usually discusses peripheral information related to the study.