Vocabulary Density Study Datasets

Document Last Updated: Wednesday, October 08, 2014

Vocabulary Density Study Datasets

Table of Contents:

Tables:

Figures:

Contact Information:

Author: James G. Mork

Email: jmork@mail.nih.gov

Phone: 301-435-3163

Vocabulary Density Study Datasets

Introduction

The inspiration for using journal-specific data to improve the NLM Medical Text Indexer (MTI) performance came from the discussion at the 2013 BioASQ Workshop[1] by Tsoumakas et al.[2] and an NLM (National Library of Medicine) senior indexer who recommended that we explore journal-specific indexing and filtering. Tsoumakas et al. used machine learning to train on only the specific journals that were involved in the BioASQ Challenge and focused on which MeSH[3] (Medical Subject Heading) Terms and how many MeSH Terms each journal typically used. To explore whether customizing the indexing for a specific journal would be worthwhile, we created the Vocabulary Density Study.

The Vocabulary Density Study summarizes how often each MeSH Descriptor and MeSH Descriptor/Qualifier combination is used by each journal in MEDLINE citations from the MEDLINE/PubMed Baseline[4]. The MEDLINE/PubMed Baseline is a snapshot created at the beginning of each new MeSH Indexing Year[5] containing the MEDLINE, OLDMEDLINE, and PubMed-not-MEDLINE completed records.

Special Notes

- In this document, we use the terms *article* and *citation* interchangeably, but they do refer to two distinct entities in the indexing world. Indexers index from the full text of an *article* and the results of that effort along with the title, abstract, and other bibliographic information from the *article* are represented by a *citation* in the MEDLINE/PubMed database.
- All indexed citations in the 2014 MEDLINE Baseline are Version 1. Although the field is available for referencing different versions of the same PMID, in this baseline, none of the indexed citations have anything other than version "1" assigned.
- MeSH Descriptors are the main descriptors or headings from the MeSH vocabulary (e.g., *Lung*) used to describe what an article is about. MeSH Descriptors can also be referred to as MeSH Terms, Main Headings, MeSH Headings, or MHs.
- MeSH Qualifiers are used to qualify the MeSH Descriptors (e.g., *Lung/abnormalities* means that the article is more about the *abnormalities* associated with the *Lung* than the *Lung* itself). MeSH Qualifiers can also be referred to as Subheadings or SHs.
- This study did not distinguish between MeSH Descriptors and Qualifiers that are the main point of an article versus those that are discussed but not considered by the indexer to be a main point.
- **Date Completed / DCOM / DateCompleted**[6]: Date Completed is the date processing of the record ends; i.e., MeSH Headings have been added, quality assurance validations are completed, and the completed record subsequently is distributed to PubMed and licensees. For records in the OLDMEDLINE subset: <DateCompleted> is the approximate date the record entered PubMed rather than of the date processing ends because OLDMEDLINE records are created and processed differently than MEDLINE records. Completion Dates are well behaved and always contain the Year, Month, and Day tags.

  XML Tag: <DateCompleted>

  XML Tags:
  <Year>
  <Month>
  <Day>

[1] http://www.bioasq.org/

[2] Tsoumakas G, Laliotis M, Markantonatos N, VlahavasI. Large-scale semantic indexing of biomedical publications at BioASQ. BioASQ Workshop, Valencia, Spain, September 27, 2013

[3] http://www.nlm.nih.gov/mesh/

[4] http://www.nlm.nih.gov/bsd/licensee/baseline.html

[5] http://www.nlm.nih.gov/bsd/mesh_indexing_date_range.html

[6] http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html#datecompleted

Vocabulary Density Study Datasets

Acronyms/Abbreviations Used

Acronym	Description
DUI	MeSH Descriptor Unique Identifier
MeSH	Medical Subject Headings
NlmID	NLM Unique Journal Identifier
QUI	MeSH Qualifier Unique Identifier
UMLS	Unified Medical Language System

Table 1 - Acronyms/Abbreviations Used in This Document

Contact Us

The 2014 Baseline Year is the first iteration of the Vocabulary Density Study datasets; please send your questions, comments, and enhancement suggestions for either the files or documentation to metamap@nlm.nih.gov.

Defining the Corpus

We used 3,392,354 citations involving 5,700 journals from the 2014 MEDLINE Baseline that have been indexed for MEDLINE over the last five years (henceforth referred to as Corpus). For each MeSH Descriptor (MH) indexed for each journal we captured the number of its occurrences (NOM) and the Number of Articles in the journal (NOA). We then normalized the frequency of each MH in the journal, computing Factor = NOM / NOA. For example, the MH *Swiss 3T3 Cells* occurred four times in the 2,231 articles of the journal *Biochemical Society (Great Britain)* in the Corpus. The Factor for this MH is 0.001793 (4 / 2231). We also performed a similar analysis capturing the MeSH Descriptor (MH)/MeSH Qualifier (SH) (e.g., *Swiss 3T3 Cells/metabolism*) combinations indexed for each journal tracking the number of their occurrences and calculating a similar Factor for the MH/SH combinations.

Several types of citations were ignored for this study, as follows:

- **OLDMEDLINE, PubMed-not-MEDLINE, any citation without a MeSH Descriptor assigned**: These types of citations were not included because we were interested only in how often MeSH Terms were used and while these all are included in the MEDLINE Baseline, they contain no indexing. The citation status is determined by lines like the following XML example: ***<MedlineCitation Owner="NLM" Status="PubMed-not-MEDLINE">***

- Any citation owned by an entity other than NLM (e.g., NASA, PIP, KIE, HMD, SIS): The indexing of citations owned by entities other than NLM may not be consistent with the indexing of citations indexed specifically for MEDLINE and identified by the *Owner=”NLM”* XML tagging. This is true even for other entities at NLM like HMD (History of Medicine Division) which have their own ownership designation (*Owner=”HMD”*). Citation ownership is determined by lines like the following XML example: ***<MedlineCitation Owner="NLM" Status="MEDLINE">***

- **Any citation with Date Completed before 2009**: We included only citations indexed during the last five years for the study to limit the drift from what MeSH Terms were used at the time the citation was indexed and changes in both MeSH and the indexing policy in the intervening time period. The MeSH vocabulary is updated every year with new terms that might be more descriptive than what was available at the time a citation was indexed, existing terms may be changed to different terms, and some terms are removed. The last two changes (updates and deletions) are handled automatically when each new Baseline is created.

- ***Comment On, Erratum For, Partial Retraction Of, Republished From,*** and ***Update Of*** Comments and Correction types determined by lines like the following XML example: ***<CommentsCorrections RefType="CommentOn">***

- ***Retracted Publication***, ***Retraction of Publication***, and ***Duplicate Publication*** Publication Types determined by lines like the following XML example: ***<PublicationType>Retracted Publication </PublicationType>***

Vocabulary Density Study Datasets

MeSH Descriptor Vocabulary Density Study Data File

MH_details_2014.txt – 331MB Uncompressed

We used the 2014 MeSH Vocabulary to identify the DUIs (column 2) and MeSH Descriptors (column 7) in this study. The example in Figure 1 shows that for *The British journal of ophthalmology* (NlmID: 0421041), the MeSH Descriptor *Eye Neoplasms* (DUI: D005134) occurs 24 times within the 1,797 articles in our Corpus. This equates to a Vocabulary Density Factor of 0.013356 (24 / 1797). The journal had a maximum of 32 MeSH Descriptors assigned to a single citation within our Corpus. The data is sorted by column 1 (NlmID).

The MeSH Descriptor Vocabulary Density Study Data file as shown in Figure 1 is a bar ("|") separated flat file containing the following fields:

1) **NlmID –** The NLM Integrated Library System unique alpha-numeric identifier for the journal
2) **MeSH DUI** – Descriptor Unique Identifier for the MeSH Descriptor
3) **DUI Frequency –** The frequency of the MeSH Descriptor within this journal in our Corpus
4) **Total Number of Articles –** Total number of indexed articles for this journal in our Corpus
5) **Vocabulary Density Factor –** Percentage calculated by DUI Frequency / Total Number of Articles
6) **Maximum Number of MeSH Descriptors –** The maximum number of MeSH Descriptors for a single citation for this journal in our Corpus
7) **MeSH Descriptor** – Preferred name of the MeSH Descriptor

Field	1	2	3	4	5	6	7
Example Row	0421041	D005134	24	1797	0.013356	32	Eye Neoplasms

Figure 1 - Example Row from MeSH Descriptor Vocabulary Density Study Data File

MeSH Descriptor/MeSH Qualifier Vocabulary Density Study Data File

MHSH_details_2014.txt – 693MB Uncompressed

We used the 2014 MeSH Vocabulary to identify the DUIs (column 2) and MeSH Descriptors/MeSH Qualifiers (column 7) in this study. The example in Figure 2 shows that for *The British journal of ophthalmology* (NlmID: 0421041), the MeSH Descriptor *Eye Neoplasms* (DUI: D005134) with the MeSH Qualifier *pathology* (Q000473) attached occurs 5 times within the 1,797 articles in our Corpus. This equates to a Vocabulary Density Factor of 0.002782 (5 / 1797). The data is sorted by column 1 (NlmID) and grouped by column 2 (MeSH DUI).

The MeSH Descriptor/MeSH Qualifier Vocabulary Density Study Data file as shown in Figure 2 is a bar ("|") separated flat file containing the following fields:

1) **NlmID –** The NLM Integrated Library System unique alpha-numeric identifier for the journal
2) **MeSH DUI** – Descriptor Unique Identifier for the MeSH Descriptor
3) **MeSH QUI** – Qualifier Unique Identifier for the MeSH Qualifier
4) **DUI/QUI Frequency** – The frequency of the MeSH Descriptor with this Qualifier within this journal in our Corpus
5) **Total Number of Articles –** Total number of indexed articles for this journal in our Corpus
6) **Vocabulary Density Factor** – Percentage calculated by DUI/QUI Frequency / Total Number of Articles
7) **MeSH Descriptor/MeSH Qualifier** – Preferred name of the MeSH Descriptor and the MeSH Qualifier attached

Field	1	2	3	4	5	6	7
Example Row	0421041	D005134	Q000473	5	1797	0.002782	Eye Neoplasms/pathology

Figure 2 - Example Row from MeSH Descriptor/MeSH Qualifier Vocabulary Density Study Data File

Additional Resources

- **MeSH Browser** – http://www.nlm.nih.gov/mesh/2014/mesh_browser/MBrowser.html
- **Introduction to MeSH** – http://www.nlm.nih.gov/mesh/intro_record_types.html
  – http://www.nlm.nih.gov/mesh/introduction.html
- **Date Completed / DCOM / DateCompleted** – http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html#datecompleted