Recent Enhancements to the NLM Medical Text Indexer

James G. Mork, Dina Demner-Fushman, Susan C. Schmidt, Alan R. Aronson

Abstract

The main goal of the US National Library of Medicine (NLM) Indexing Initiative is to explore indexing methodologies that may help the NLM Indexing staff keep pace with the ever increasing challenges of indexing over 700,000 MEDLINE citations each year using a vocabulary of over 27,000 MeSH Descriptors and 220,000 MeSH Supplementary Concept Records. The BioASQ Challenge has been a tremendous benefit by expanding our knowledge of other indexing systems, specifically the technologies used in those systems to identify relevant indexing for biomedical literature. This paper provides an update on improvements to NLM's Medical Text Indexer (MTI) functionality and performance since the first BioASQ Challenge. We have, in a limited way, applied some of the lessons learned from that first Challenge to MTI to assess what performance gains we might see. The research discussed at the 2013 BioASQ Challenge Workshop inspired us to make changes to MTI that have resulted in a 2.69 (4.44%) increase in Precision and very little change in Recall.

Enhancements

- **Vocabulary Density**

  On average, only 999 unique MHs of the 27,149 available in 2014 MeSH are used per journal in the 6,606 journals in our Corpus. 83.81% of the used MHs are found in 500 or fewer journals and 271 MHs are only found in a single journal (see Figure 1). This selective use of MHs confirms the intuition that taking into account journal-specific data can lead to improvements in MTI recommendations. **Furthermore, implementing this simple approach leads to a 2.69 (4.44%) improvement in Precision, 1.36 (2.23%) increase in F_1 score, and a 0.05 (0.08%) increase in Recall.**

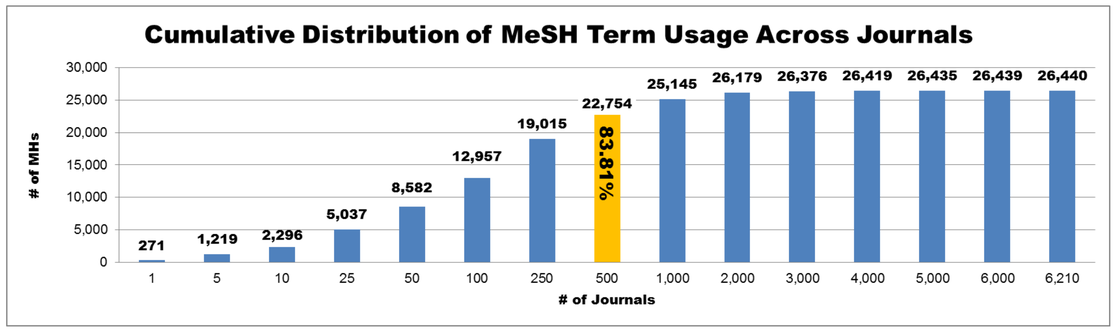

Figure 1. Cumulative Distribution of MeSH Term Usage Across Journals

- **Out of the Ballpark (OOTB) Term Filtering**

  For example, if the article is about a *3-arm clinical trial* and MTI recommends *Arm*. *Arm* would be considered an OOTB term since it is completely unrelated to any of the final indexing. Ambiguity is the primary cause of why MTI recommends an OOTB.

  - ✓ **Metaphorical Ambiguity** - *Birds of a Feather Working Group* triggering MTI Recommendations *Birds* and *Feathers*
  - ✓ **Brand Name Ambiguity** - *commit murder* triggering MTI Recommendation *Tobacco Use Cessation Products* because *Commit* is a brand name
  - ✓ **Psychology Term Ambiguity** - *employee retention* triggering MTI Recommendation *Retention (Psychology)*
  - ✓ **Body Part/Disease Tree Filtering**- *Ankle joint* triggering MTI Recommendation *Ankle*, but, the article discusses "sprained ankles" triggering indexing of *Ankle Injuries*. These are not completely unrelated, but, still marked as OOTB since *Ankle (A01.378.610.250.149)* comes from a different MeSH Tree than *Ankle Injuries (C26.558.100)*.

  Through OOTB filtering, we were able to remove 3,175 OOTB terms (10.92%) with only a 0.94 loss in Recall. More importantly, we discovered most of the OOTB terms do not represent egregious errors like *Arm* in our *3-arm clinical trial* example.

- **Preliminary New Work**

  We have also computed the Vocabulary Density for every Journal/MeSH Heading/MeSH Subheading combination and are seeing **improvements of up to 42.39 (97.77 %)** in our Subheading Attachment (SAP) Precision. **The overall Precision has gone up 12.08 (21.25%).** This focus on improving Precision does come at a cost of 14.22 (41.99%) to Recall.

Main References

1. James G Mork, Antonio J Jimeno-Yepes, Alan R Aronson. The NLM Medical Text Indexer system for indexing biomedical literature. BioASQ Workshop, Valencia, Spain, September 27, 2013.
2. Tsoumakas G, Laliotis M, Markantonatos N, VlahavasI. Large-scale semantic indexing of biomedical publications at BioASQ. BioASQ Workshop, Valencia, Spain, September 27, 2013.