Proceedings of AMIA Annual Symposium 1999;:1127.

Automated Assignment of Medical Subject Headings

Stuart J. Nelson, MD, Alan R. Aronson, PhD, Tamas E. Doszkocs, PhD, W. John Wilbur, MD, PhD, Olivier Bodenreider, MD, PhD, H. Florence Chang, MS, James Mork, MS, and Alexa T. McCray, PhD
National Library of Medicine, Bethesda, MD

Introduction. As part of the National Library of Medicine's Indexing Initiative, we developed and compared automated methods of assigning Medical Subject Headings (MeSH) to MEDLINE citations.

Methods. A test collection of 200 MEDLINE citations published in 1997, with abstracts in English, were selected at random. The following methods of finding and ranking suitable MeSH descriptors have been investigated using this test collection:
The Inquery Algorithm. This algorithm depends on parsing text into noun phrases, then using the Inquery search engine[1] to match to MeSH descriptors. Co-occurring MeSH descriptors in the UMLS are used to suggest additional headings.
MetaMap. MetaMap[2] develops an ordered list of UMLS Metathesaurus concepts for each citation, based on the noun phrases extracted from that text. A ranked list of concepts is developed for each phrase.
Trigram Algorithm. A phrase is broken into overlapping trigrams (three letters occurring in succession) for analysis. Candidate phrases are obtained from the title and abstract by examining all maximal contiguous sets of words that contain no punctuation or stop words (from a list of 310 common stop words). The trigrams are used to match phrases in the UMLS, with the maximal overlap of sets of trigrams resulting in the suggested UMLS concept.
Restricting to MeSH. Once a UMLS concept has been identified, using the MetaMap method or the Trigram algorithm, the task becomes one of navigating within the UMLS to find the appropriate MeSH heading. This method was described previously[3].
Related Articles Method. This method depends on the assumption that the semantic neighbors of a document are those documents in the database that are the most similar to it[4]. The similarity between documents is measured by the words they have in common, with some adjustment for document lengths. The test document is used as the basis for finding similar documents. MeSH descriptors assigned to similar documents are then assigned to the test document.
Clustering and Weighting of Suggested Headings. After using one or more of the above methods, the suggested MeSH headings are clustered. Descriptors close together in the same MeSH trees are given additional weight, as are the descriptors known to co-occur with high frequency in MEDLINE. The suggested headings from each method being tested are then presented in rank order.

Results and Conclusions. A formal trial of the methods has not yet been completed. However, several observations can be made. Use of different parsers to extract noun phrases from title and abstract did not appear to significantly alter the performance. In clustering and weighting the suggested headings, the most important aspect appeared to be the number of times a given descriptor was suggested. Second in importance was the semantic relationships between descriptors. The numerical value of the weighting factors had little effect. It appears that methods based on natural language processing and mapping to MeSH are complementary to the related articles method, and that any system should therefore use a combination of those methods.

References

1. Callan JP, Croft WB, Harding SM. The INQUERY Retrieval System. *Proceedings of the 3rd International Conference on Database and Expert Systems Applications* 78-83, 1992.

2. Aronson AR, Rindflesch TC, and Browne AC. Exploiting a large thesaurus for information retrieval. *Proceedings of RIAO 94*, 197-216, 1994.

3. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond Synonymy: Exploiting the UMLS Semantics in Mapping Vocabularies. *J Am Med Informatics Assoc* (Symposium Suppl), 815-9, 1998.

4. Wilbur WJ, Coffee L. The effectiveness of document neighboring in search enhancement. *Information Processing & Management*, 30(2):253-266, 1994.