Use of the Metathesaurus and SPECIALIST Lexicon of the Unified Medical Language System, Lexical Matching and Domain-Specific Free-Text to Identify Undocumented Vocabulary

Eric M. Mills D.V.M. Ph.D., Jeff R. Wilcke, D.V.M., M.S., Diplomate, A.C.V.C.P., Holly S. Bender, D.V.M., Ph.D. Diplomate, A.C.V.P.
Virginia-Maryland Regional College of Veterinary Medicine

ABSTRACT

This technique uses lexical matching within domain-specific free-text to locate and remove terms contained in the Metathesaurus or SPECIALIST Lexicon. The remaining character strings of text were presented to domain experts along with the original sections of text for manual review. Internet search engines were used to verify questionable character strings. The key element of this technique was the SPECIALIST Lexicon which was used to remove common words and known variations of medical terms from the text. The removal of these terms reduced the number of non-numeric character strings remaining in the text from 12,075 to 374. Ninety-three percent of 230 randomly selected remaining character strings resulted in the identification of undocumented vocabulary terms.

METHODS

The textbook, Medical History and Physical Examination in Companion Animals (1) was used with kind permission from Kluwer Academic Publishers. The Index and chapters relating to the canine physical examination were scanned, converted to electronic text and spellchecked for errors using a custom veterinary dictionary created by parsing the CD-ROM version of the Merck Veterinary Manual (2) into a list of unique words. Next, the electronic text was converted to a table, one record for each sentence, phrase or heading.

English terms contained in the Metathesaurus and the SPECIALIST Lexicon of the 1997 release of the Unified Medical Language System (UMLS) were imported from the UMLS distribution files into tables. A subset of uniquely spelled terms (case-insensitive) from each table was extracted and converted to uppercase with classification or notation characters removed. (e.g., “Abdomen, NOS” and “Cold ≤1”) Terms from each table were sorted in order of decreasing word count. A word was defined as any alphabetic text string delimited by spaces and/or punctuation.

Lexical matching (case-insensitive) was performed in two sequential phases, 1) using terms contained in the Metathesaurus followed by 2) using terms contained in the SPECIALIST lexicon. Matched terms in either phase were replaced with a meaningless token character. The remaining character strings were exported to a table, one record for each unique string, along with its frequency of occurrence.

RESULTS

Phase one searching resulted in 12,224 remaining unique character strings, with 12,075 containing at least one alpha character. Phase two searching resulted in 806 remaining unique character strings, with 574 containing at least one alpha character. Of the 574 character strings, 213 (.93) of 230 randomly selected strings were useful in identifying undocumented vocabulary terms (medical or non-medical). Evaluation was performed by (EMM) with review by (JRW) to reach consensus. Character strings which were not immediately apparent as being useful were submitted for searching on the Internet using the Metacrawler search service.

DISCUSSION

The key element of this technique is the SPECIALIST Lexicon used to remove common words and known variations of medical terms from the text. Manual analysis of the character strings that remained after phase two searching revealed complete terms, partial terms, proper names, errors in the original text, scanning errors and various combinations of the above. This technique will not identify multiword terms for which all words have been documented previously in the Metathesaurus or SPECIALIST Lexicon.

References