Journal Descriptor Indexing Tool for Categorizing Text according to Discipline or Semantic Type

Susanne M. Humphrey[1], Chris J. Lu, Ph.D.[2], Willie J. Rogers[2], Allen C. Browne[1]
[1] National Library of Medicine, Bethesda, Maryland 20894
[2] Management Systems Designers, Fairfax, Virginia 22031

Abstract. A JDI (Journal Descriptor Indexing) tool has been developed at NLM® that automatically categorizes biomedical text as input, returning a ranked list, with scores between 0-1, of either JDs (Journal Descriptors, corresponding to biomedical disciplines) or STs (UMLS® Semantic Types). Possible applications include WSD (Word Sense Disambiguation) and retrieval according to discipline. The Lexical Systems Group plans to distribute an open source JAVA version of this tool.

NLM (National Library of Medicine®) maintains two broad, relatively small classifications:

- A set of 122 descriptors from MeSH® (Medical Subject Headings®), known as JDs, used for manually indexing MEDLINE® journals *per se* according to discipline. These are found in the List of Journals Indexed for MEDLINE, which also contains the listing of titles under these descriptors. For example, Journal of Pediatric Surgery is listed under both Pediatrics and Surgery.
- A set of 135 STs in the Semantic Network in NLM's UMLS (Unified Medical Language System®). Concepts in the UMLS Metathesaurus® are assigned one or more STs which semantically characterize those concepts. For example, the Metathesaurus concept Aspirin is assigned the STs Pharmacologic Substance and Organic Chemical.

The JDI tool uses a methodology based on statistical word-JD associations from a training set of MEDLINE citations to which are imported the JDs corresponding to journal unique identifiers in the citations. For example, words in articles in the Journal of Pediatric Surgery become statistically associated with the JDs Pediatrics and Surgery. Then an input text comprised of words similar to the ones in these articles would be categorized by the same JDs. Using words in the input, JDI ranks the JDs according to the average of JD scores in word-JD associations. For example, the first three JDs, with scores, returned by JDI for the input "appendectomy in children" are: 1 0.7311 Surgery, 2 0.6856 Pediatrics, and 3 0.4661 Gastroenterology.

This methodology is being applied in the SemRep UMLS NLP (Natural Language Processing) tool; JDI increases accuracy by identifying MEDLINE citations in the molecular genetics domain before NLP begins. A possible retrieval application would be to intersect citations described by a JD with citations described by textwords or another JD, for example: neurotransmitters [tw] AND Cardiology [jd]; Cardiology [jd] AND Pediatrics [jd].

JDI methodology is the basis for STI (Semantic Type Indexing). ST "documents" are created comprised of UMLS Metathesaurus strings belonging to the ST, and these documents each undergo JDI. Then statistical word-ST associations are calculated by comparing JDI of individual training set words and JDI of these ST documents. Using words in the input, STI ranks the STs according to the average of ST scores in word-ST associations. For example, the first three STs, with scores, returned by STI for the input "appendectomy in children" are: 1 0.5985 Age Group, 2 0.5520 Finding, and 3 0.5498 Therapeutic or Preventive Procedure. That is, the average Age Group score for words in the input is higher than for other STs. An alternate method of STI compares the JDI of the input to the JDI of each ST document, and ranks the STs according to the greatest similarity to their ST documents. By this method, JDI of this input is most similar to JDI of the Age Group document.

NLM has applied STI to WSD. If the senses of an ambiguous word are expressed by candidate STs for its meaning, STI can be performed on the context surrounding the word (phrase, sentence, abstract) in the expectation that in the STI of the context, the correct ST for the word will rank higher than the other candidate STs.

The Lexical Systems Group plans to distribute an open source JAVA version of the JDI tool[*] as part of the UMLS NLP tools. This tool would allow users to enter text input, and would return a ranked list of JDs or STs with scores between 0-1.

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

[*] For more details and to download the open source tool, please visit the Text Categorization web site at *http://specialist.nlm.nih.gov/tc*