Implementing Comprehensive Derivational Features in Lexical Tools Using a Systematical Approach

Chris J. Lu, Ph.D.[1, 2], Destinee Tormey[1], Lynn McCreedy, Ph.D.[1] and Allen C. Browne [1]
[1]National Library of Medicine, Bethesda, MD; [2]Medical Science & Computing, Inc, Rockville, MD

Abstract
A systematic approach for automatically generating derivational variants based on the SPECIALIST Lexicon was proposed and implemented in Lexical Tools [1]. This approach addressed the prefix (PD), zero (ZD), and suffix (SD) derivations from nominalizations (nomD). This paper describes the generation of SD (not from nomD) based on the Lexicon in the Lexical Tools, including both SD-Facts and SD-Rules. New derivational features, such as negation, derivation types, and enhanced algorithms are also included in the Lexical Tools 2013 release for better precision and recall in NLP applications.

1. Introduction
The Lexical Tools can be used to retrieve lexical variations, including derivations. Query expansion by substituting subterms with derivations (closely related terms that may differ in syntactic category) is an effective Natural Language Processing (NLP) technique for better recall without dropping precision, yielding a better result. For example, if the source vocabulary includes uricosuric|adj (not in UMLS), the derivational flow (-f:d) will map it to uricosuria|noun, which is a UMLS Metathesaurus term. More information, such as concepts (C0151582) and synonyms, can be retrieved for further NLP analysis.

2. SD-Rules and SD-Facts
The 97 SD-Rules collected in the Lexical Tools are used as candidate rules for generating SD-pairs. First, 42,089 SD-pairs matching these candidate rules are retrieved from the Lexicon. About 28% (12,087) of them are from nomD and tagged as valid (relevant) by computer. The rest (72%) are sent to linguists for manual tagging. The tagging results of these SD-Rules are sorted by the descending order of precision (= relevant, retrieved No./retrieved No.) and then retrieved No. The system performance of a subset can be calculated from the sum of cumulative (cum.) precision and recall (= relevant, retrieved No./ relevant No.) of rules included in the subset by the above sorting order. An optimized set with the top 65 rules from the 97 candidate rules is obtained by: 1) removing 10 duplicated child rules 2) evaluating 9 related parent rules 3) finding the intersection of curves of system precision and recall. This model is also used to evaluate new SD-Rules. 10 new SD-Rules (4 from nomD, 5 from factD, and 1 from suggestion) are evaluated. Finally, we obtain an optimized set with the cutoff at the intersection of the two curves of precision and recall, including 73 SD-rules (out of 96 unique candidate rules) with 95.30% system precision and 95.01% system recall, as shown in Figure-1 [2].

The valid SD-Pairs generated from all candidate SD-Rules are added to SD-Facts. The coverage of SD-Facts is dramatically increased to an order of magnitude (from 4,559 in 2012 to 44,832 in 2013) with this approach. The restriction algorithm of SD (-kd) in Lexical Tools is enhanced (SD-Facts based) to filter out invalid dPairs in the Lexicon to reach virtually 100% accuracy (assuming no tagging error). The optimized set of SD-Rules is used to predict derivations not in the Lexicon and is expected to have above 95% precision.

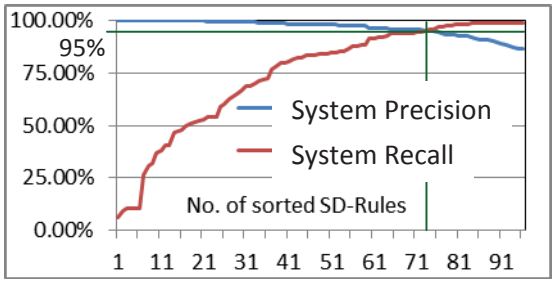

Figure-1. Performance of Optimized Set of SD-Rules

3. Other New Features and Conclusion
Two new derivational options are implemented in Lexical Tools: 1) -kn:N|O|B to specify negation, otherwise, or both. By default (-kn:O), negative derivations are filtered out because the shift/opposite meaning is usually not desired. For example, "anti-convulsive" (a negation of "convulsive") is filtered out when the concept of "convulsive" is the only focus. Negations can be retrieved by –kn:N if opposite meaning is desired. 2) -kt:P|S|Z to specify the derivation type(s) of PD, SD, ZD, or any combinations. Visit http://specialist.nlm.nih.gov/lvg for more details. This approach provides a maintainable and scalable system for generating derivations with the Lexicon's annual release and results in comprehensive derivational features in the Lexical Tools for better precision and recall. Lexical Tools is distributed by NLM via an Open Source License agreement.

Acknowledgments
This research was supported by the Intramural Research Program of the NIH, National Library of Medicine. The authors would like to thank Mr. Guy Divita, Mr. Howard Lu, and Dr. Fiona M. Callaghan for their valuable discussions and suggestions.

References
1. C.J. Lu, L. McCreedy, D. Tormey, and A.C. Browne., "A Systematic Approach for Automatically Generating Derivational Variants in Lexical Tools Based on the SPECIALIST Lexicon", IEEE IT Professional Magazine, May/June, 2012, p. 36-42
2. http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lvg/current/docs/designDoc/UDF/derivations/SD-Rules-Opti/index.html