Classification Types: A New Feature in the SPECIALIST Lexicon

Chris J. Lu, PhD[1, 2], Amanda Payne, PhD[1, 2] and Dina Demner-Fushman, MD, PhD [1]
[1]National Library of Medicine, Bethesda, MD [2]Medical Science & Computing, LLC, Rockville, MD

Introduction

The SPECIALIST Lexicon (thereafter, the Lexicon) [1], distributed by the National Library of Medicine (NLM) as one of the Unified Medical Language System (UMLS) knowledge sources, supports popular NLP tools, such as SemRep, MetaMap, cTAKES, CSpell, and the SPECIALIST Lexical Tools, as an underlying resource. A new enhanced feature called the classification type (CT) is a proposed addition to the Lexicon. These classification types can be archaic, source, informal, or other. First, terms classified as archaic, such as *cozen*, *colde* and *benight*, are considered no longer in common use in modern corpora (such as MEDLINE). These terms may have modern equivalents in the same lexical record (*colde* for *cold*) or in separate ones (*ye* for *the*). Second, normalization on spelling variants from foreign English into US English is needed if the source is from a foreign country. For example, British English (*analyse*, *leukaemia*, *tumour*) can be normalized to US English (*analyze, leukemia, tumor*). These terms are classified as source. Third, consumers often use informal language when they ask questions. For example, *bomb* for *success*, or *grandpa* for *grandfather* are used primarily in colloquial contexts. The performance of automated consumer question understanding could be improved if the Lexicon provides informal terms with their cross-referenced (CR) formal terms (synonyms). Four CTs and their syntax are shown in Table 1.

Table 1. Enhanced Features of Four Classification Types.

Code	Examples/Description
class_type=archaic|base form	class_type=archaic|colde
class_type=source|base form|originated language	class_type=source|analyse|british
class_type=informal|base form|CR-citation|CR-EUI	class_type=informal|bomb|success|E0058772
class_type=other	other type of classification (i.e. gene, protein, etc.)

Classification Type Implementation

CTs are added to new lexical records during lexicon building by NLM linguists through a Web-based tool, LexBuild. New GUI components for adding CTs and the enhanced LexCheck software for validating syntax and contents of CTs are integrated in LexBuild. CT tagging on existing lexical records is also done through LexBuild. First, class_type=unassigned is added to all existing records, and removed after it is tagged. Computer-aided features for retrieving records by specified patterns, like suffix, substring, category, etc. are implemented for systematic tagging. Finally, post-process programs are implemented to generate new Lexicon tables, including archaic terms, spelling variants with originated sources and informal terms with their formal synonyms, for the Lexicon annual release.

Applications and Conclusion

The performance of NLP applications that use the Lexicon is expected to be improved with the new CTs. First, archaic terms can be excluded when dealing with modern corpora. Second, foreign English can be normalized to US English with source CTs. Third, the performance of automated question understanding on consumer health can be improved. For example, *grandpa* (no concept found in UMLS) can be effectively mapped to *grandfather* (C0337475) with query expansion by substituting formal synonyms for their informal terms [2]. The Lexicon is distributed by NLM via an Open Source License agreement and is available at: https://umlslex.nlm.nih.gov/lexicon.

Acknowledgements

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health. We thank Dr. L. McCreedy, D. Tormey and A.C. Browne for their valuable discussions.

References

1. https://umlslex.nlm.nih.gov/lexicon
2. Lu CJ, Tormey D, McCreedy L, Browne AC, Enhanced LexSynonym Acquisition for Effective UMLS Concept Mapping, MedInfo 2017, Hangzhou, China, August 21-25, 2017, 245:501:505