Using Element Words to Generate (Multi)words for the SPECIALIST Lexicon

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
The SPECIALIST Lexicon has been distributed annually by the National Library of Medicine (NLM) since 1994. Lexical records are used for Part-of-Speech (POS) tagging, indexing, information retrieval, concept mapping, etc. in many Natural Language Processing (NLP) projects, such as Lexical Tools, MetaMap, SemRep, UMLS Metathesaurus, and ClinicalTrials.gov. This paper describes a new systematic approach to identify single words and multiwords from MEDLINE through the use of element words. Element words are lowercase single words without punctuation and are not stopwords. Results show an accelerated growth of the Lexicon, particularly an increase in multiword records. Hence, improvement in recall or precision can be anticipated in NLP projects using the SPECIALIST Lexicon and its applications.

1. Introduction – The NLP SPECIALIST Lexicon and LexBuild

The Lexicon is built by linguists through a web-based computer-aided tool, LexBuild [1]. Element words are a resource used by linguists to 1) add new Lexical records if no exact/close match is found in LexBuild; 2) update existing lexical records if related records are found by close match. Multiwords that contain these new element words are reviewed through the Essie search engine [2], Google Scholar, dictionaries, etc. during the LexBuild process.

2. (Multi)words by New Element Words from MEDLINE

For (multi)word inclusion in the Lexicon from MEDLINE, we developed this system: 1) retrieve element words through tokenization (lowercase, remove punctuation, and use space as word boundaries) from MEDLINE titles and abstracts; 2) categorize these element words by type: single words in the Lexicon (e.g. diabetes), not a single word but parts of multiwords already in the Lexicon (e.g. mellitus), numbers (e.g. five), digits (e.g. 5), non-words (e.g. 3h), and new element words (e.g. cdh); 3) calculate word count (WC). New element words with high frequency (WC >= 1500) are retrieved automatically for review to cover single words (97.58%) and multiwords from MEDLINE. For example, the new element word "cdh" (9983 WC) leads to 44 new lexical records with base forms in 78 single words (e.g. cadherin1) and 23 multiwords (e.g. "chronic daily headache").

3. (Multi)words by Existing Element Words from MEDLINE

This system also retrieves candidates of new multiwords from MEDLINE for (existing) element words: 1) Generate high frequency n-grams of length 1-5 from MEDLINE. The low frequency n-gram terms are filtered out if the associated (n-1)-gram terms have low WC of normalized form (NWC). 2). N-grams are normalized by abstracting away from genitive, punctuation, and case so that different forms of a same term are grouped together for further analysis. 3). Generate new candidate multiwords by applying a rule-based system to filter out invalid multiwords from n-grams. These rules exclude (normalized) n-grams that exist in the Lexicon, start/end with a preposition/auxiliary/modal/conjunction, end with determiner/acronym in a parenthesis, etc.. Document count, WC, and NWC are also used to filter out low frequent error prone n-grams (e.g. typos). Further development of these rules is intended to increase the precision of candidate multiwords. 4). These new candidate multiwords are reviewed by linguists, who add grammatical and lexical variant information, yielding completed Lexicon records. For example, the element word "mellitus" was identified in 24 multiword lexical records in the previous Lexicon release (2014). In our new approach, a candidate list (532) is retrieved automatically from 1304 n-gram terms containing "mellitus" after filtering out ~60% of invalid words. This list is then mapped into 390 normalized forms to ease the final review process in linguistic contexts. As a result, 36 new lexical records with base forms in 9 single words (including, actually, "mellitus") and 41 multiwords (e.g. "diabetic mellitus") have been added, a 150% growth from Lexicon.2014. In addition, 7 other associated existing records with 10 multiwords have been updated for spelling variants and acronym expansions. Please refer to the Lexicon web site for details [3].

4. Conclusion

There are 477K lexical records with 1.69M forms in the 2014 Lexicon release. About 47.7% (418K) of unique forms (875K) are multiwords. Multiwords are an essential ingredient and play a key role in the success of NLP tasks. This new system enhances the Lexicon's coverage, especially on multiwords. We expect the growth of multiwords in future Lexicon releases to reach the estimated value (50%) through this system [4]. This new system encourages rapid growth of single words as well as multiwords in the Lexicon, which will ultimately provide better NLP results.

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. N.C. Ide, R.F. Loane, D.D. Fushman, "Essie: A Concept-based Search Engine for Structured Biomedical Text", JAMIA, Vol. 14, Num. 3, May/June, 2007, p.253-263
3. http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lexicon/2015/docs/designDoc/UDF/medline/index.html
4. I.A. Sag, T. Baldwin, F. Bond, A. Copestake, D. Flickinger, "Multiword Expressions: A Pain in the Neck for NLP", Computational Linguistics and Intelligent Text Proc., Lecture Notes in Computer Science, Vol. 2276, 2002, p. 1-15