Enhancing LexSynonym Features in the Lexical Tools

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

Introduction

Concept mapping is vital to natural language processing (NLP) for bioinformatics. Query expansion using synonyms for subterm substitutions is an effective technique to increase recall when no direct concept mapping can be found through normalization. For example, no concept can be found by direct mapping through normalization if the source vocabulary is "calcaneal fracture". By substituting the subterm "calcaneal" for its synonym, "heel bone", the matched UMLS concept [C0281926, fracture of calcaneus] is found. In this example, "calcaneal" and "heel bone" are element synonyms while "heel bone fracture" is the expanded term. The effectiveness of this technique relies on the completeness and quality of both the element synonyms and the UMLS synonym thesaurus in the query expansion pipeline of UMLS concept mapping (Figure 1).

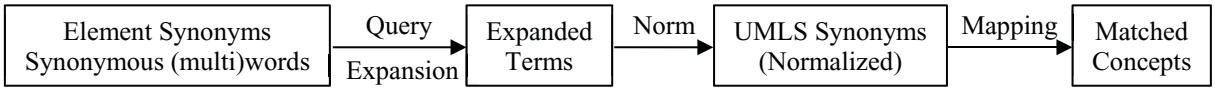

Figure 1. Query Expansion Pipeline of UMLS Metathesaurus Concept Mapping

Enhanced LexSynonym Features in the Lexical Tools

A systematic approach to LexSynonym acquisition from the UMLS Metathesaurus and the Lexicon was developed to meet two necessary properties, commutativity and transitivity, for quality element synonyms [1]. LexSynonyms are synonymous (multi)words in the Lexicon, such as "calcaneal" and "heel bone" in the above example. 190,844 LexSynonyms are acquired in the Lexicon.2017. They are integrated in the Lexical Tools (2017) for synonym retrieval. The synonym flow component (-f:y) is enhanced to include synonyms, POS and source information. A sophisticated algorithm is implemented in the recursive synonym flow component (-f:r) to preserve precision based on the source types. In addition, the synonym source option (-ks) is added to restrict the results by source type (CUI, EUI, NLP), or any combination of the above. These new features provide needed information for downstream NLP processing. For example, the size of the fruitful variants flow component (-f:v, using recursive synonym features), that is used in various NLP projects (Sophia, MMTx, Custom Taxonomy Builder) for better recall, has been increased from 8.9M (2016) to 9.9M (2017) due to the enhancement of LexSynonyms.

Test, Results and Conclusion

The UMLS-CORE project assigned CUI(s) to terms (13,076) that are within the top 95% usage and mappable to SNOMED CT [2]. 2,755 of these terms (with 2,756 CUIs) that do not have direct mapped concepts in UMLS.2016AB are used to test the performance of LexSynonyms for query expansion. Three different synonym sets are tested through the Sub-Term Mapping Tools (STMT) [3]. The results show improvements in recall (4.97%) and F1 (0.05) from both the 2016 and 2017 LexSynonyms with the STMT synonym set (Table 1) [1]. Precision is slightly increased (0.26%) due to the high quality of LexSynonym 2017 (71.04% precision in this test). The set of LexSynonyms and enhanced features are distributed in the Lexical Tools.2017 with UMLS by NLM via an Open Source License agreement. Improvements in performance can be anticipated for NLP applications that use these enhanced features of LexSynonyms.

Table 1. Test Result for LexSynonyms 2016-17

Synonym Set	TP*	FP*	Rel.*	Precision	Recall	F1
STMT + 2016	691	358	2,756	65.87%	25.07%	0.3632
STMT + 2017	828	424	2,756	66.13%	30.04%	0.4132
2017	287	117	2,756	71.04%	10.41%	0.1816

*T: True, F: False, P: Positive, Rel.: Relevant

References

1. CJ Lu, D Tormey, L McCreedy, AC Browne, Enhanced LexSynonym Acquisition for Effective UMLS Concept Mapping, MedInfo 2017, HangZhou, China, Aug. 21-25, 2017, accepted for publication.
2. KW Fung and J Xu, An exploration of the properties of the CORE problem list subset and how it facilitates the implementation of SNOMED CT, JAMIA 2015, 22: 649-658.
3. CJ Lu and AC Browne, Development of Sub-Term Mapping Tools (STMT). In Proceedings of AMIA 2012 Annual Symposium, Chicago, USA, Nov. 3-7, 2012, p. 1845.