Enhancing The SPECIALIST Lexicon with WordNet

Chris J. Lu, PhD[1, 2], Amanda Payne, PhD[1, 2] and James G. Mork, MSc[1]
[1]National Library of Medicine, Bethesda, MD [2]Medical Science & Computing, LLC, Rockville, MD

Introduction

WordNet® is a large lexical database of English [1]. It is used to enhance the SPECIALIST Lexicon (the Lexicon) for expanding coverage of (multi)words, derivations, synonyms and antonyms [2]. Computer programs are developed to retrieve candidates of words, derivation pairs (dPairs), synonym pairs (sPairs) and antonym pairs (aPairs) from WordNet. The results show high performance (F1) on these candidates. This paper describes our systematic approach and primary results on zero dPairs.

Methods

Computer programs were developed to retrieve candidates of lexemes and dPairs from WordNet for the Lexicon as follows [3]: 1) retrieved dPairs in WordNet 3.0 by utilizing the Java WordNet Interface (JWI) [4]; 2) standardized and categorized dPairs into zero, prefix and suffix dPairs; 3) generated zero dPair candidates by filtering out zero dPairs with invalid patterns; Zero dPairs are derivations with the same spelling and different Parts of Speech (POS). For example, smart|noun is a zeroD of smart|verb, but an invalid (zero) derivation of smart|adjective. 4) generated word candidates from zero dPairs by filtering out words with invalid patterns; 5) applied UMLS Metathesaurus CUI (2021AB) and MEDLINE n-grams (2021) as filters to improve performance; 6) sent these dPairs and word candidates to linguists for annotation and addition to the Lexicon; 7) analyzed results.

Results, Discussion and Future Work

The results show: 1) About 84.07% of 4,751 zero dPairs derived from WordNet were tagged in the Lexicon (2022). These tagged zero dPairs had high precision (98.25%) as valid dPairs. Untagged zero dPairs (757) are used as dPair candidates to be added to the Lexicon and expected to have high precision. 2) Words from zeroD pairs, which are not tagged in the Lexicon and matched valid word patterns of combined filters, are used as word candidates (1,368) for Lexicon building. This word candidate list has a high precision (86.11%). 3) The performance could be further improved by applying CUI, MEDLINE or a combination of both filters. As shown in Table 1, the MEDLINE filter has the best overall performance of F1 (88.46%), while the CUI filter has slightly higher precision (90.15%) with much lower recall (30.31%) and results in low F1. This indicates that MEDLINE covers most valid words and is a good filter for lexeme retrieval. Applying both CUI and MEDLINE filters reaches the highest precision (90.44%).

Table 1. Performance of word candidates from WordNet zero dPairs on CUI, MEDLINE and CUI & MEDLINE filters.

Filter(s)	Precision	Recall	F1
CUI	90.15%	30.31%	45.36%
MEDLINE	88.12%	88.79%	88.46%
CUI & MEDLINE	90.44%	29.71%	44.73%

We plan to apply similar approaches to retrieve candidates of suffix dPairs, sPairs, aPairs and words from WordNet to effectively build the Lexicon and enrich coverage. The Lexicon is distributed with Unified Medical Language System (UMLS) by National Library of Medicine (NLM) via an Open Source License agreement and is available at: https://lhncbc.nlm.nih.gov/LSG/Projects/lexicon/current/.

Acknowledgements

This research was supported by the Lister Hill National Centre for Biomedical Communications (LHNCBC) of the NLM, National Institutes of Health.

References

1. Miller GA. WordNet: A Lexical Database for English. Communications of the ACM, 1995, Vol. 38, No. 11: 39-41
2. Lu CJ, Payne A, Mork JG. The Unified Medical Language System SPECIALIST Lexicon and Lexical Tools: Development and applications. JAMIA, May 29, 2020, Vol. 27, Issue 10, Oct. 2020, pp: 1600-1605.
3. https://lhncbc.nlm.nih.gov/LSG/Projects/lexicon/current/docs/designDoc
4. Finlayson, M A. Java Libraries for Accessing the Princeton Wordnet: Comparison and Evaluation. In H. Orav, C. Fellbaum, & P. Vossen (Eds.), Proceedings of the 7th International Global WordNet Conference (GWC 2014), pp: 78-85. Tartu, Estonia.