Identifying Criteria for Antonym Generation from Corpora

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

The SPECIALIST Lexicon (the Lexicon) and Lexical Tools are enhanced with new antonym features for the 2022 release [1]. Antonym pairs (aPairs) are generated from the Lexicon, suffix and prefix derivations, collocates (CC) and semantic relations (SC) in corpora [2]. Criteria are identified from a set of common aPairs to effectively generate antonyms from CC and SC. This paper describes our systematic approach on identifying generic criteria for antonym generation from corpora.

Methods

We collected commonly used antonyms from 14 sources on the internet [3]. This collection includes 1000 unique, lowercased, single word aPairs, which are assumed to have representative characteristics of overall antonyms and is used as a training and test set (TtSet) to identify generic criteria of antonyms. APairs in the TtSet are manually tagged for canonicity, domain, type, and negation. Canonical aPairs have a generic domain, that is central to human life and ways of living across time and cultures. Computer programs are developed to 1) retrieve properties of aPairs, such as EUIs, POSs, CUIs, STIs, sources, etc. 2) compute stats among properties to identify generic criteria of antonyms. TtSet and 2021 antonym production data are used for this study.

Results – Identified Criteria

First, aPairs from TtSet and 2021's data set have the same (10) domains for canonical antonyms and similar negation rates (9.51% and 7.14%, respectively). This conforms with our hypothesis that TtSet is representative of overall antonyms to retrieve antonym characteristics. Second, Table 1 shows the source distribution of TtSet. We observed antonyms from corpora models (CC and SC) to be worthy of further development because they contain the most aPairs (83.65%). Third, canonical aPairs must be in the Lexicon with valid EUIs and have the same POS. Fourth, canonical aPairs cannot be synonyms. This confirms the theory that antonyms and synonyms are similar in domain and different in polarity. Fifth, antonyms should have CUIs (our scope is using concepts in the UMLS-Metathesaurus) and share STIs because 67.79% of aPairs share STIs when they have CUIs. The analysis' results are shown in Table 2 and used to retrieve antonym candidates from corpora. Our ultimate objective is to provide generic and comprehensive antonym features with completion of antonym generation from corpora models. The Lexicon is distributed with UMLS by NLM via an Open Source License agreement and is available at: https://umlslex.nlm.nih.gov/lexicon.

Table 1. Source distribution of canonical aPairs in the TtSet.

	LEXICON	SuffixD	PrefixD	CC	SC
TtSet canonical aPairs	1.95%	0.58%	13.81%	33.07%	50.58

Table 2. Stats of analyzed properties for canonical aPairs in the TtSet.

	EUIs	Same POS	Synonyms	Share STIs (if have CUIs)
TtSet canonical aPairs	100%	100%	0%	67.79%

Acknowledgements

This research was carried out by staff of the National Library of Medicine (NLM), National Institutes of Health, with support from NLM.

References

1. 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, p: 1600-1605.
2. Lu CJ, Payne A, Mork JG. Enhanced Features in the SPECIALIST Lexicon - Antonyms. AMIA, 2020 Virtual Annual Symposium, Nov. 14-18, 2020. p: 1833.
3. https://lhncbc.nlm.nih.gov/LSG/Projects/lexicon/current/docs/designDoc/UDF/antonyms/2-4.AntSource-TT.pdf