4. WordNet – Antonyms (APairs)

1. Antonyms form WordNet
   - From WordNet:
     - Directory: ${WORDNET}
     - File: ${WORDNET}/data/output/**WnAPairs.data.3.0**
     - **WnAPairs.data.3.0**: 12,248 aPairs
     - format: base1|base2|pos
     - These are antonym pairs from WordNet directly.
     - Antonyms must have the same POS
     - Antonyms include single words and multiwords
   - **Task 4.1: Generate aPair candidates from WordNet aPairs.**
     - Directory: ${LEXICON_ANTONYM}/../WordNetAPairs
     - Input file: WnAPairs.data.3.0
     - Output File: ${LEXICON_ANTONYM}/data/6.WordNet/2021/outData/WordNet/Cand
     - Program: GenAPairCand.java
     - Used Lexicon.2021 for the initial baseline data (Log.WnApCand.2021)

Steps	Description	Notes
1	Unify sPair and remove duplicates	• WordNet aPairs are duplicated and are not in a standard format. They are unified and standardized in this step. • WnAPairs.unique.data (8,242)
2	Remove Lexicon SpVar	• WordNet aPairs include spVars. SpVars are not aPairs and need to be removed. • Wn.aPairCand.data.trap.spVar (2)
3	Apply combined filters	• Words in aPairs must be valid words. Use combined filter to remove invalid words. • wn.aPairCand.data.trap.cf (107)
4	Filter out Lexicon sPairs (2022)	• Remove Lexicon sPairs. • wn.aPairCand.data.trap.sp (1)
5	Baseline output (sorted)	• wn.aPairCand.data.baseline (8,132)
6	Filter out Lexicon tagged aPairs	• wn.aPairCand.data.yes (964) • wn.aPairCand.data.no (380) • wn.aPairCand.data.tbd (6,788)

     - The baseline candidate file (used for paper) use Lexicon.2021. This is used to calculate precision for paper.
     - The final candidate file sent to linguist should use the latest Lexicon, by re-running the program to get the latest results.
     - The procedures are:
       - Send all word candidates from Task below (Task 4.2) to linguist for LexBuilding.

        - Update inflVars (inflVars.data.current) and notBaseWord.data once above step is done
        - Go through annual aPairs process to update aPairs.
        - Re-run this step with latest Lexicon and notBaseWord.data, The tbd should become zero because invalid aPairs are excluded from candidates or in the tagged file and valid aPairs are in the tagged file.
    - For the baseline, there are 1344 (= 964+380) among 8,132 WordNet aPair candidates (16.53% = 1344/8132) are overlapped). The precision is estimate as 71.13% (=964/(964+380)). This implies there are much more aPairs could be retrieved from the WordNet, it is a good resource for Lexicon aPair retrieval.
  - WordNet includes 2 pairs of spVars as aPair:
    - kern|kern|verb
    - ravel|ravel|verb
  - WordNet includes 1 sPair as aPair:
    - mushroom|toadstool|noun

- **Task 4.2: Generate word candidates from aPairs in WordNet (To be updated)**
  - Directory: ${LEXICON_ANTONYM}/../6.WordNet
  - Input file: ${LEXICON_ANTONYM}/data/6.WordNet/2021/outData/Cand/wn.aPairCand.data.2021
  - Output File: ${LEXICON_SYNONYM}/data/6.WordNet/2021/outData/Cand/wn.Ap.wordCand
  - Program: GenWordCandFromAPairs.java
  - Used Lexicon.2021 for the initial baseline data (Log.WnApWordCand.2021)

Steps	Description	Notes
0	Retrieve words from sPairs	Get unique words from aPairs (9,904)
1	Combined filters • Standard filter • Single word filters • Multiword filters	• Apply combined filter to filter out invalid word • wn.Ap.wordCand.1.cfTrap (7,595)
2	WordNet Previous cand filter • verb complement • zeroD • suffixD	• Exclude words from previous WordNet word candidates • wn.Ap.wordCand.2.pcTrap (373)
3	MEDLINE n-gram filter	• Use both MEDLINE 1-gram and n-grams (WC >=30) • wn.Ap.wordCand.3.mlTrap (695)
4	UMLS CUI filter	• Use UMLS CUI • wn.Ap.wordCand.4.cuiTrap (1,355)
	Output	• **wn.Ap.wordCand.cand.2021 (1,936)** • wn.Ap.wordCand.cuiPass (581, have CUIs)

- Use Lexicon.2021 for the baseline in the paper.
- Use Lexicon.current for practice and send to linguists.
- Combined filters: Lexicon, general, pattern, single-word, multiword-lead-end-term filters.
- Previous Candidate filters: not in the previous candidate list in the WordNet model (verb-complement, zeroD and suffixD)
- MEDLINE filter: use both 1-gram for single words and n-gram (wc >= 30) for multiwords
- CUI filter: must have UMLS CUIs
- Output: **wn.Ap.wordCand.cand**
    - Use Lexicon current (recompile) and WordNet.aPair.2021 to generate this file.
  - Send to linguists for lexBuilding (**TBD**)
    - No need to tag
    - Analyze the precision on WC and ML and CUI
  - Format: word|POS|WC|Tags
  - Includes words in/not in MEDLINE, with/without CUIs, maybe use a cutoff WC for initial test (**wn.Ap.wordCand.cand.wc10000, wc > 10000)**
  - Tags:
    - TAG_ML_1G_CUI_NO, TAG_ML_1G_CUI_YES
    - TAG_ML_5G_CUI_NO, TAG_ML_5G_CUI_YES
    - TAG_NOT_ML_CUI_NO, TAG_NOT_ML_CUI_YES

======= What we used before this new Model (To BE Deleted) ==========

- The antonym criteria (STI, CUI, Single word, EUI, etc., 693 aPairs) are not used here.

- If not use STI, CUI criteria, it extended to 8,252 aPairs (can be used in the future)
  - Check if in aPair tagged file (216 aPairs from previous tag, 693-216=477): 477 aPairs.
  - Convert to aPair candidate format (include ants without EUIs): 477 aPairs.
  - Shell> GetAntonyms 2021

  - Output: ${ANTONYM}/data/6.WordNet/outpt/Cand/APairCand.data
  - Tagged: [N]:155, [Y]: 215, [TBD]: 107, total 370.