Antonyms Analysis – Training and Test Set (TtSet)

The collected antonyms from the training and test set (TtSet) are assumed to have representative characteristics of the overall antonyms and are used to identify generic properties of antonym pairs (aPairs). APairs in the TtSet are manually tagged for canonical, domains, types, and negations. Computer programs are developed to 1) retrieve properties of these aPairs, such as EUIs, POSs, CUIs, STIs, sources, etc. 2) compute stats among properties to identify generic criteria of antonyms. These criteria include properties of EUI (Entry unique identifier), POS (Part-Of-Speech), concepts (CUIs – Concept Unique Identifier), semantic type (STI – Semantic Type Identifier) and synonyms. The identified criteria are then implemented in the antonym generation model to find antonym candidates from CC (collocates in MEDLINE). They are discussed as follows.

1. Retrieve Properties for Antonyms

Computer programs are developed to retrieve properties for all aPairs from TtSet. They are described below.

1. Algorithm: GetProperties.java to check if
    - Ant1 and ant2 are single words
    - Ant1 and ant2 have EUIs (in the Lexicon): [EUI_Y|EUI_1|EUI_2|EUIN]
    - Ant1 and ant2 have the same POS when both must have EUI: [POS_Y|POS_N]
    - Ant1 and ant2 have CUIs (valid concept in UMLS): [CUI_Y|CUI_1|CUI_2|CUI_N]
    - Ant1 and ant2 share same STI (Semantic type), when both must have CUI: [STI_Y|STI_N]
    - Ant 1 and ant2 are synonyms to each other: [SYNONYM_Y|SYNONYM_N]

Where

- EUI:
    - [EUI_1]: only ant1 has EUI
    - [EUI_2]: only ant2 has EUI
    - [EUI_N]: neither ant1 or ant 2 has EUI
    - [EUI_Y]: both ant1 and ant2 have EUI
- POS:
    - [POS_N]: ant 1 and ant2 does not have the same POS
    - [POS_Y]: ant 1 and ant2 have the same POS
- CUI:
    - [CUI_1]: only ant1 has CUI
    - [CUI_2]: only ant2 has CUI
    - [CUI_N]: neither ant1 or ant 2 has CUI
    - [CUI_Y]: both ant1 and ant2 have CUI
- STI:
    - [STI_N]: ant 1 and ant2 does not share same STI
    - [STI_Y]: ant 1 and ant2 share same STI
- SYNONYM:
    - [SYNONYM_N]: ant1 and ant2 are not synonyms
    - [SYNONYM_Y]: ant1 and ant2 are synonyms

2. Program:

```
Shell> cd  ${ANTONYM}/bin
Shell> GetAntonyms ${YEAR}
45
```

1.3 Output – format and examples:

Output file is ./analysis/antonymTtSet.data.properties. It has 8 fields (please refer to output file for details). Table 1 shows an example of the output file from antonym candidates derived from Ttset.

Ant1	Ant2	Source	EUI	POS	CUI	STI	Synonym
with	without	LEX	EUI_Y	POS_Y	CUI_2	STI_N	SYNONYM_N
always	never	LEX	EUI_Y	POS_Y	CUI_Y	STI_Y	SYNONYM_N
accept	refuse	CC	EUI_Y	POS_Y	CUI_Y	STI_Y	SYNONYM_N
start	stop	CC	EUI_Y	POS_Y	CUI_Y	STI_Y	SYNONYM_N
admit	deny	SN	EUI_Y	POS_Y	CUI_Y	STI_Y	SYNONYM_N
adore	hate	SN	EUI_Y	POS_Y	CUI_Y	STI_Y	SYNONYM_N
able	unable	PD	EUI_Y	POS_Y	CUI_Y	STI_Y	SYNONYM_N
possible	impossible	PD	EUI_Y	POS_Y	CUI_Y	STI_Y	SYNONYM_N
careful	careless	SD	EUI_Y	POS_Y	CUI_N	STI_N	SYNONYM_N
worthy	worthless	SD	EUI_Y	POS_Y	CUI_2	STI_N	SYNONYM_N
…	…	…	…	…	…	…	…

Table 1. Examples of properties of antonym collection from TtSet.

2. Get Stats on properties of antonyms

A program is developed to calculate the stats among properties in the previous section. This program is run on two data sets of: 1). 1000 aPairs from TT; 2). 514 canonical aPairs from TtSet.

1. Algorithm: GetAntPropertyStats.java
    - Input file: ./analysis/antonymTtSet.data.properties
    - Calculate percentage on the following properties:
        - Source, EUI, POS, CUI, STI, SYNONYM
        - POS under EUI
        - STI under CUI

2. Program

```
Shell> cd ${ANTONYM}/bin
Shell> GetAntonyms ${YEAR}
46|47 (each option for one of the two data sets)
50|51|52 (split, then on training set)
```

3. Output file and results:
   The output stats file are:
   - ./analysis/antonymTtSet.data.pStats
   - ./analysis/antCandTtSet.data.tag.Y.pStats

   The results and analysis from the above two files are described below:

3.1. Results:

APairs from TtSet that are not from source of [LEX|SD|PD] are temperately assigned as source from [TT]. These aPairs are then checked with MEDLINE n-gram set to retag the source as 1) [CC] - collocates in MEDLINE or 2) [SN] – not collocates in MEDLINE, namely they are semantical antonyms in corpora. There are two possible for aPairs with source of [SN]:

1). They are collocates in other corpus, but no collocates in MEDLINE. For examples,

- seller|buyer: “seller market and buyer market” can be found in other corpus;
- compliment|insult: is collocates in iWeb corpus (https://www.collocates.info/iweb.asp)

2). They are not collocates in any corpus. For example:

- abominate|love might not be in any corpus because abominate is such a rare word, so it is possible some of these are just not relevant for the collocate model.

Table 2 shows:

1) among the most commonly used 1000 aPairs (candidates) collected from TtSet, over 90.40% are from CC (32.20%) and SN (58.20%)
2) among the 514 canonical aPairs (tagged) from TtSet, over 83.66% are from CC (33.07%) and SN (50.58%)

A summary of analyses are described below based on the observation of the results from this program.

Analysis-1:

Source of CC (collocates in MEDLINE) contains about 1/3 distribution for both antonym candidates (32.20%) and canonical antonyms (33.07%). Currently, we have completed model development for antonym generation from source of LEX|SD|PD and antonym candidates from PD are still under tagging (tagging is completed for LEX and SD). It is imperative to develop antonym generation model from CC and other models from SN to provide a comprehensive antonym coverage.

	Total	**LEX**	**SuffixD**	**PrefixD**	**CC**	**SN**
TtSet (candidates)	1000	10 (1.00%)	7 (0.70%)	79 (7.90%)	322 (32.20%)	582 (58.2%)
TtSet (canonical)	514	10% (1.95%)	3 (0.58%)	71 (13.81%)	170 (33.07%)	260 (50.58%)

Table 2. Source distribution of the antonym collection from TtSet

3.2. EUI (in the Lexicon)

Table 3 shows 99.90% and 100% of antonyms are in the Lexicon for antonym candidates and canonical antonyms, respectively.

Analysis-2:

Antonyms must be in the Lexicon.

	Total	**None**	**Ant1**	**Ant2**	**Both**
TtSet (candidates)	1000	0 (0.00%)	1 (0.10%)	0 (0.00%)	999 (99.90%)
TtSet (canonical)	514	0 (0.00%)	0 (0.00%)	0 (0.00%)	514 (100.00%)

Table 3. EUI distribution of the antonym collection from TtSet.

3.3. POS (have the same POS)

Table 4 shows:

1) Among the most commonly used 1000 aPairs (candidates) collected from TtSet, over 97.50% have the same POS.

2) Among the most commonly used 1000 aPairs (candidates) collected from TtSet, over 97.60% have the same POS if antonyms are in the Lexicon (have EUIs).

3) Among 514 canonical aPairs (tagged) from TtSet, 100% have the same POS.

Analysis-3:

Antonyms in aPairs must have the same Part-Of-Speech (POS).

	Total	**No**	**Yes**
TtSet	1000	25 (2.50%)	975 (97.50%)
TtSet (Both have EUIs)	999	24 (2.40%)	975 (97.60%)
TtSet (canon)	514	0	514 (100.00%)

Table 4. Distribution of the antonym collection from TtSet with the same POS.

3.4. CUI (Have UMLS concepts)

Table 5 shows:

Among the most commonly used 1000 aPairs and canonical aPairs collected from TtSet, in only about 51.95% ~ 55.18% of them, both antonyms have CUIs. However, our research scope is using concepts in the UMLS-Metathesaurus. Thus, our requirements are set as antonyms must have valid CUI.

Analysis-4:

Our aPairs are a more strictly defined (smaller) set than generally used antonyms. This is appropriate because we are targeting precision when applying antonyms in the NLP applications. We can't find any concept for further NLP process anyway for those antonyms without CUIs.

	Total	None	Ant1	Ant2	Both
TtSet	1000	138 (13.80%)	170 (17.00%)	147 (14.70%)	545 (54.50%)
TtSet (with same POS)	975	132 (13.54%)	163 (16.72%)	142 (14.56%)	538 (55.18%)
TtSet (canonical)	514	91 (17.70%)	90 (17.51%)	66 (12.84%)	267 (51.95%)

Table 5. CUI distribution of the antonym collection from TtSet.

3.5. STI (Share same Semantic Types)

Table 6 shows:

1) Among the most commonly used 1000 aPairs collected from TtSet, about 32.00% of them share same semantic types.
2) Among canonical aPairs (tagged) from TtSet, over 67.79% of antonyms share same semantic types if they both have CUIs.
3) Among canonical aPairs (tagged) from TtSet, over 69.10% of antonyms share same semantic types if they both have CUIs and the source is CC or SN.

Analysis-5:

Applying semantic type criteria on aPairs reduces the antonym candidates to a smaller and higher precision set than commonly used antonyms. It is appropriate for targeting higher precision NLP applications (the tradeoff is dropping the recall).

	Total	Not share STI	Share STI
TtSet	1000	680 (68.00%)	320 (32.00%)
TtSet (canonical, both have CUIs)	267	86 (32.21%)	181 (67.79%)
TtSet (canonical, both have CUIs, CC|SN)	228	75 (32.89%)	153 (69.10%)

Table 6. Same STI distribution of the antonym collection from TtSet.

2.3.6 Synonym (is synonym)

Table 7 shows none of the antonyms are synonyms.

Analysis-6:

Antonyms cannot be synonyms. This confirms the theory that antonyms and synonyms are similar in domain and different in polarity.

	Total	No	Yes
TtSet	1000	1000 (100.00%)	0 (0.00%)
TtSet (canon)	514	514 (100.00%)	0 (0.00%)

Table 7. Synonym distribution of the antonym collection from TtSet.

2.3.7 Domains

Table 8 shows 10 domains and tagged examples found in canonical aPairs from TtSet.

No.	Domain	Tagged Examples
1	existence	birth|E0013159|death|E0020918|noun|Y|UB|BN2|existence|CC
2	frequency	always|E0008403|never|E0042565|adv|Y|UB|N2|frequency|LEX
3	location	ceiling|E0015728|floor|E0028200|noun|Y|UB|O|location|CC
4	physical_property	visible|E0064742|invisible|E0035728|adj|Y|B|O|physical_property|PD
5	possibility	admit|E0007437|deny|E0021749|verb|Y|B|BN2|possibility|CC
6	quality	careful|E0015340|careless|E0015344|adj|Y|UB|O|quality|SD
7	quantity	all|E0008090|none|E0042838|pron|Y|UB|N2|quantity|CC
8	size	dwarf|E0024153|giant|E0029703|noun|Y|UB|O|size|CC
9	temperature	cool|E0018931|warm|E0065055|verb|Y|UB|O|temperature|CC
10	temporal	early|E0024315|late|E0036937|adv|Y|UB|O|temporal|CC

Table 8. Domains used in the antonym collection from TtSet.

3. Get stats from the tagged antonym candidates

The collected antonyms from TtSet are used to generate antonym candidates as described in document, 2-4.AntSource-TT. A program is developed to find the stats for the tagged antonym candidates from TtSet. This same program is generic and applied to all tagged antonym candidates to generate stats as well. The latest antonym generation data from 2021 are used as a subset to represent the overall antonyms in this analysis.

1. Algorithm: GetStatsFromTagCand.java
    - Input file: ./output/candTagged/antCandTtSet.data.tag/tagged/${YEAR}
    - Calculate percentage on the following:
        - Canon & source
        - Source
        - Canon & POS
        - POS

    - Type
    - Canon & negation
    - Domain

2. Program

```
Shell> cd ${ANTONYM}/bin
Shell> GetAntonyms ${YEAR}
48|5|6
```

3. Output files and results:
   The output stats files are:
   1) option – 48: TtSet: ./output/analysis/antCandTtSet.data.tag.stats
   2) option – 5: 2021 Data (all tagged aPairs): ./output/analysis/antCand.data.tag.stats
   3) option – 6: 2021 Data (Canonical): ./output/analysis/antCand.data.tag.Y.stats

The results and analysis of the above three files are described below:

3.1. Domains

The 2021 tagged data includes antonyms generated from LEX|SD|PD and contains 3558 antonym candidates, and have the same 10 domains as found in TtSet (sec 2.3.7). Please note that antonym candidates are completely tagged for LEX and SD, while PD is currently partially tagged.

Analysis-7:
Having the same 10 domains corresponds to our hypothesis of using TtSet as a representative set for overall antonyms.

3.2. Canonical aPairs

APairs from TtSet and 2021 are used to compare the canonical rate in the antonym candidates. First, the 1000 antonyms are expanded to 1252 antonym candidates by expanding antonyms with their spelling variants. Only 45.77% among these antonym candidates are tagged as canonical antonyms, as shown in Table 9. The canonical rate for 2021 antonym candidates (55.71%) is higher than the most commonly used antonyms (TtSet). This implies our antonym generation model is effective to generate antonym candidates. The canonical rate of antonym candidates from 2021 data will be more accurate and have more meaning once PD is completely tagged.

Analysis-8:
Not all commonly used antonyms are legit aPairs suitable for high precision NLP applications. Only canonical antonyms are chosen in our study. With derived criteria, we expect to generate higher precision antonym candidates for effective antonym generation.

	Antonym Candidates	Canonical	Not-canonical
TtSet Candidates	1252	574 (45.77%)	679 (54.23%)
2021 Candidates	3558	1982 (55.71%)	1576 (44.29%)

Table 9. Canonical distribution of antonym candidates in TtSet and 2021 Data.

3.3. Estimated overall canonical aPairs by source of SD

Analysis-9:

Currently, we have completely tagged and generated aPairs from the source of suffix derivation (SD) for 2021 data. The number of canonical aPairs from SD is 132. Thus, we estimated the total canonical aPairs is 22,758 (= 132/0.0058), the percentage of SD is 0.58% from Table 2. Please note that we do not use the tagged canonical aPairs from LEX for the estimation because that number is rather static and does not grow with the growth of corpora (the Lexicon). We will also estimate the total canonical aPairs by PD once PD is completely tagged to confirm our estimation.

3.4. POS distribution

Table 10 shows the POS distribution for canonical aPairs from TtSet and 2021 data. The 2021 data is uncompleted (PD is not completely tagged), so the distribution is not 100% representative. However, the top four POS (Adj, Noun, Verb and Adv) for canonical aPairs are the same. Please note that canonical aPairs from the rest of the POSs (Modal, Pron, Aux, Prep, Det and Conj) are rather static and were retrieved from the Lexicon (because they are associated with negation tags in the Lexicon).

POS	**TtSet (Canon)**	**2021 Data (Canon)**
Adj	42.06%	66.75%
Noun	26.35%	15.89%
Verb	22.16%	12.82%
Adv	5.76%	2.06%
Modal	0.35%	0.66%
Pron	1.22%	0.56%
Aux	0.00%	0.51%
Prep	1.39%	0.40%
Det	0.35%	0.20%
Conj	0.35%	0.15%

Table 10. POS distribution of canonical aPairs in TtSet and 2021 Data.

Analysis-10:

The top four POS distribution of canonical aPairs between TtSet and 2021 data is the same. This corresponds to our hypothesis of using TtSet as a representative set for overall antonyms.

3.5. Negation distribution

Table 11 shows the negation distribution for antonym candidates from TtSet and 2021 data. Both sets have similar negative and not-negative rate. Please note that negation is independent from the canonical property of an aPair. Accordingly, the negation distribution from candidates (including noth canonical and not canonical aPairs) are used for bigger sampling coverage.

Analysis-11:
The distribution of negation rate of aPairs between TtSet and 2021 Data are similar. This corresponds to our hypothesis of using TtSet as a representative set for overall antonyms.

POS	**TtSet (Cand)**	**2021 Data (Cand)**
True Negative	1.76%	1.69%
Broadly negative	7.75%	5.45%
Not-negative	90.50%	92.86%

Table 11. Negation distribution of antonym candidates in TtSet and 2021 Data.