Antonyms Evaluation – Training and Test Set (TtSet)

We use TtSet to train and test criteria in antonym generating models. Precision, recall and F1 are used as metrics to measure the performance. Retrieved instances are aPairs derived from antonym generating models, while relevant instances are canonical aPairs annotated by linguists. They are described as follows.

1. **Derived criteria from training set and evaluation on test sets**

   There are 1000 aPairs collected in the TtSet. The TtSet are split randomly into 80% training and 20% test set (process-52). As a result, the training and test set include 799 and 201 aPair instances, respectively. We analyzed three properties of EUI, POS and synonyms on the training set (process-53). The result shows:

   1). All antonyms are in the Lexicon, that is all antonyms in the 799 aPairs have EUIs.

   2). 97.87% antonyms have the same POS in the 799 aPairs.

   3). None (0.00%) of antonyms are synonyms. This confirms the theory that antonyms and synonyms are similar in domain and different in polarity.

   These three criteria are then evaluated on the test set. APairs were retrieved by 4 different criteria, EUI, POS, not synonym and combination of above three criteria in this evaluation (process-54). The results show precision and F1 were increased, and the recall is preserved by applying these three criteria, as shown in Table 1. We concluded these three criteria are valid and the combination of them should be used in antonym generating models.

Criteria	Precision	Recall	F1
None	0.5124	1.0000	0.6776
1. Must have EUI	0.5150	1.0000	0.6799
2. Must have same POS	0.5337	1.0000	0.6959
3. Must not be synonyms	0.5124	1.0000	0.6776
Combination of 1, 2 & 3	0.5337	1.0000	0.6959

Table 1. Results for criteria of EUI, POS, and synonyms on the TtSet

2. **Evaluation on instances with UMLS CUI**

The scope of the antonym generation task is using concepts in the UMLS-Metathesaurus because the Lexicon is one of the three major components to support NLP research using UMLS. Accordingly, one of the requirements is to limit antonyms to have valid CUIs. There are 545 aPairs which have CUIs in the TtSet. We conducted the same evaluation as above and the result is shown in Table 2 (process-55). The results confirm these three criteria improve precision and F1 while preserving recall for the scope of our task (antonyms have CUIs).

Criteria	Precision	Recall	F1
None	0.4899	1.0000	0.6576
1. Must have EUI	0.4908	1.0000	0.6584

2. Must have same POS	0.4963	1.0000	0.6634
3. Must not be synonyms	0.4899	1.0000	0.6576
Combination of 1, 2 & 3	0.4963	1.0000	0.6634

Table 2. Results for criteria of EUI, POS, and synonyms on aPairs with CUIs on the TtSet.

3. Evaluation on instances with UMLS CUI on CC sources

Our goal is to find criteria to apply to the CC (collocates in corpus) model to improve performance. Thus, we shifted our focus on the instances that are from CC source in the TtSet. There are 271 aPairs that have CUIS and are derived from CC in the TtSet. This set is used to evaluate criteria for the CC model.

We added a new criteria that aPairs must have same STI (semantic type). These four criteria were evaluated and the result is shown in Table 3 (process-56). This new criteria of having same STI increases the precision, yet drops the recall and F1 (by 0.02). In practice, we applied all 4 criteria to generate antonym candidates from CC model to increase the precision.

Criteria	Precision	Recall	F1
None	0.5129	1.0000	0.6780
1. Must have EUI	0.5148	1.0000	0.6797
2. Must have same POS	0.5187	1.0000	0.6830
3. Must not be synonyms	0.5129	1.0000	0.6580
4. Must same STI	0.5497	0.7554	0.6364
Combination of 1, 2, 3 & 4	0.5556	1.7554	0.6402

Table 3. Results for criteria of EUI, POS, synonyms and STI on aPairs with CUIs and CC in the TtSet.

4. Evaluation of collocates with n-gram models

Antonyms are often collocates in corpora. This phenomenon is used to retrieve antonym candidates from a selected corpus in the collocates in corpora model (CC). The MEDLINE net gram set were used as a corpus for the collocates model. Antonym collocates appear in 3-grams, 4-grams and 5 grams because antonyms must be a single word. Table 4 shows examples of antonyms in the 3-grams, 4-grams and 5-grams (process-57). A performance evaluation on N-grams (N= 3 ~ 5) was conducted on the TtSet because aPairs from TtSet appear as collocates in different N-grams (process-58), as shown in Table 5. In general, the collocate instances of 5-grams are a subset of 4-grams; and the collocate instances of 4-grams are a subset of 3-grams. The MEDLINE 3-grams were chosen as the corpus in CC model for the best recall and F1 performance.

Our antonym generation model includes source from 1) Lexical records with negative tag (LEX); 2) suffix derivations with negation (SD); 3) prefix derivations with negation (PD); 4) collocates from corpus (CC). The CC model applies the MEDLINE 3-grams to have better recall. In our test, there are

582 aPairs (58.2%) in the TtSet that are not retrieved from our model. It is imperative to develop models to have a comprehensive coverage for antonym generation.

TBD: we could use existing antonym corpus (wordnet) to retrieve antonym candidates:

aPair	increase|decrease	alive|dead	copy|original
3-grams	• 5934|increase or decrease • 1990|increase and decrease • 940|decrease or increase • 691|decrease and increase • 205|decrease with increase • …	• 218|dead or alive • 198|alive or dead • 94|alive and dead • 45|dead and alive • …	None
4-grams	• 1662|increase or decrease in • 965|increase or decrease the • 775|an increase or decrease • 693|increase and decrease in • 662|to increase or decrease • …	None	• 417|copy of the original
5-grams	• 528|an increase or decrease in • 439|increase or decrease in the • 342|an increase or a decrease • 291|decrease with an increase in • 277|increase or a decrease in • …	None	• 387|copy of the original print • 387|scanned copy of the original

Table 4. Antonym example in MEDLINE 3-grams, 4-grams, 5-grams.

	precision	recall	F1
3-grams	0.6029	0.4922	0.5419
4-grams	0.6301	0.4258	0.5082
5-grams	0.6544	0.3477	0.4547

Table 5. Performance on TtSet for using MEDLINE N-gram (N= 3~5) in CC model.