Interim Report - LexAntonyms

Hi Jim,

Amanda and I have been working on the antonym task since the end of March. We have developed some processes with good primary results. This R & D task includes implementation, annotation and documentation. The general processes for antonym generation include:

- Developing programs to generate antonym candidate lists.
- Annotate the antonym candidates.

  The annotation is a complicated process because each antonym pair (aPair) involves multiple tags on [CANONICAL ANTONYM], [TYPE], [NEGATION] and [DOMAIN].
- Developing programs to automatically check annotation conflicts and fix tagged files.
- Developing programs to generate aPairs and negation detection cue words (NDCW) from verified tagged files.

There are four sources/models used to generate antonyms. The status and accomplishments are described as follows.

1). Lexical records (LEX) with negative and broad negative tags:

We completed adding both negative and broad negative tags into the Lexicon through LexBuild for the POSs of adverb, pronoun, auxiliary, modal, preposition, determiner and conjunction. We also developed programs to generate aPairs (45) and NDCWs (44) using the Lexicon 2020 release. This sub-task is estimated with 95% completion.

2). Suffix derivation (SD) with negative tags:

We completed the development of computer programs to accomplish the general processes of antonym generation from SD. Antonym candidates from SD were generated and annotated. APairs (128) were generated from the tagged antonym candidates from the Lexicon 2020 release. No NDCW was found from the source of SD. This sub-task is estimated with 90% completion.

3). Collocates from a Corpus (CC):

This sub-task is the most challenging one. The approach is described as follows:

- We plan to develop a systematic methodology that can be applied to different corpora to retrieve antonyms.
- MEDLINE was chosen as the corpus to retrieve aPairs for the first attempt (due to the availability of MEDLINE n-gram set).
- Established a Training and Test set (TtSet) of aPairs from 14 antonym web sites. This TtSet includes 1,255 commonly seen unique aPairs.

- Annotating the TtSet based on the definition in our antonym document (Amanda is currently working on the annotation)
- Developed algorithm to derive patterns by applying co-occurring and using all 75% verified aPairs (TtSet). This program was developed on MEDLINE 3-grams with temporary data set.
- Developed program to generate antonym candidates from the derived patterns using MEDLINE n-gram set. The program was developed and tested on MEDLINE 3-grams.
- Calculate the recall (compare to 25% of verified TtSet) and precision (based on tagging) on the generated candidates. A good result could lead us to a publication (TBD)
- To develop programs to check, auto-fix and generate aPairs and NDCWs (TBD).

This sub-task is estimated with 60% completion for MEDLINE 3-grams. More patterns should be developed if time permitted.

4). Prefix derivation (PD) with negative tags:

The development of this sub-task should be similar to the development of SD. I expect there are much more antonym candidates in this categories. Our plan is to start the development once we complete the sub-task of CC.

The antonym task is a big task (similar to synonyms) and probably needs 2-3 years to complete for sources of CC and PD with current linguists resources. I will continue to work on this task when needed. Hope this makes sense. Please let me know if you have any questions. Thank you!