Updating UMLS Metathesaurus hierarchy b-tree files:

These hierarchy b-tree files are used to determine if a hypernymic ISA relation holds between two UMLS Metathesaurus concepts. We need flat text files that hold this information and create Berkeley DB b-tree files from them[1]. Due to Linux limitation on file size and to the fact that SemRep is 32-bit, we split the data into smaller chunks[2], so that they can be accessed by SemRep. An appropriately-sized chunk contains 80M records.

SemRep configuration file (*SAWenv*) contains two environment variables that are relevant to the location of the btree files: *UMLS_HIERARCHY_DIR* corresponds to the directory where the b-trees reside and *UMLS_HIERARCHY_DB_PREFIX* corresponds to the b-tree file prefix, to which UMLS version number and the file number are added to form the actual b-tree hierarchy file names. The current directory (*UMLS_HIERARCHY_DIR*) is *${DATA_DIR}/UMLS_HIERARCHY*[3] and the prefix is *${UMLS_HIERARCHY_DIR}/hrel-UMLS_btree.*, indicating that the b-tree files have the form *hrel-UMLS_btree. 2006AA_X* where *X* is the file number and that these files are located in *${DATA_DIR}/UMLS_HIERARCHY* directory.

Creating the b-tree files is somewhat complicated, due to incompatible versions of Berkeley DB[4] and the procedure is summarized here. First, the script that creates the b-tree files from a text file is a Perl script (*create_hrel_btree.pl* in *$SAW/UMLSTransitiveClosure*). To be able to use Berkeley DB routines in Perl, Perl BerkeleyDB module has to be installed. This installation is not default in Perl installation. There are other complicating issues:

- DB files created using the default Perl interpreter on Linux machines (perl 5.8.8), for some reason, did not work with SemRep. To overcome this, I installed perl-5.8.9 locally (*/rhome/kilicogluh/perl-5.8.9*).

- The current Berkeley DB version required by SemRep is 4.8.24 (since MetaMap uses this version). There is an installation of this version of Berkeley DB at *$NLS/tools/Berkeley_db/Linux-i686* directory, but Perl module is not installed, so we cannot use this installation directly. I created a separate version of Berkeley DB-4.8.24 locally, by copying *$NLS/tools/Berkeley_db/Linux-i686/db-4.8.24.NC* directory locally, and building a 32-bit db-4.8.24 from it. I took the following steps in db-4.8.24.NC directory.

    - *cd build_unix*

[1] The current flat text files are provided by Medical Ontology Research group (Olivier Bodenreider) and reside in */nfsvol/crfiler-semrep/DATA/UMLS_HIERARCHY* directory (i.e., *${DATA_DIR}/UMLS_HIERARCHY*).

[2] This was the case for 2006AA version, but MOR group changed their algorithm to create this file recently and the resulting file now seems to fit in one b-tree, so some of the details in this file may not be relevant anymore, but are kept here for reference.

[3] *${DATA_DIR}* refers to SemRep data file directory, currently */nfsvol/crfiler-semrep/DATA*.

[4] These incompatibility issues might be resolved in future releases of RH Linux and Berkeley DB, therefore, it is sensible to try to create b-tree files with the existing versions before going forward with the rest of the steps.

  - CONFIGURE: *../dist/configure –enable-java –prefix=/rhome/halil/BerkeleyDB/db-4.8.24_32 –enable-static –build=i686-pc-linux-gnu.*
  - MAKE, TEST, INSTALL: *make, make test, make install*[5]

- Next step is to enable Perl module to work with the Perl installation.
  - *cd db-4.8.24.NC/perl/BerkeleyDB*
  - CONFIGURE: Update *config.in* file to point to *$prefix/include* and $prefix/lib directories.
  - MAKE, TEST, INSTALL: *perl-5.8.9/perl Makefile.PL, make, make test, make install*.
  - This associates the BerkeleyDB module with the Perl installation.
- The shell script *create_hrels.sh* in *${SAW}/UMLSTransitiveClosure* is used to create the database files.
  - *./create_hrels.sh <TEXT_FILE_DIR> <BTREE_DIR> <NUMBER_OF_CHUNKS>*
    - Some assumptions are made regarding the file names and the location of the Perl installation. The script should be updated, if necessary.
    - *create_hrel_btree.pl* is required to be in the same directory (*$SAW/UMLSTransitiveClosure*)
- After creating b-tree files, they should be tested. The test script is called *test_hrel_btree.pl* (*$SAW/UMLSTransitiveClosure*). An example way of testing is below:
  - *tail -10f <TEXT_FILE> | perl test_hrel_btree.pl -bt <CORRESPONDING_BTREE_FILE>*

[5] *make install* copies to the directory indicated by the prefix config option.