Integrating SKR into SemRep: [1]

SemRep uses a slightly modified version of the official SKR/MetaMap program (which we call *semSKR*). The modifications consist of several SemRep-specific changes to SKR (e.g., dysonym processing) as well as addition or modification of predicates used by SemRep. This file details these differences and the procedure of integrating SKR into SemRep. The procedure may need to be updated for subsequent versions of SKR or SemRep.

My current strategy for this integration is to rename the existing SKR used by SemRep, copy the new SKR source code and make the changes needed for SemRep on the new SKR source code. This is at this time manual and is error-prone, future work might involve automating this to some extent. The problem is that SKR/MetaMap source code changes independently of SemRep and some of the changes that affect SemRep are not necessarily easy to find out. So, there is a bit of trial/error and debugging in integrating SKR into SemRep.

Three directories of SKR are relevant to SemRep: *bin*, *prolog* and *src*.

***bin* directory:**

This directory contains scripts that set up SKR environment as well as compile and debug SKR.

1. *SKRenv*[2]*:* Sets up SKR environment, similar to *SAWenv* in *$SAW/bin*. Both files overlap significantly and the settings that are shared are likely to (and, for the most part, should) be the same. Particular settings that may need to be changed include *SKR*, *SAW*, *SKR_SRC_HOME*, and *SAW_SRC_HOME* which set up SKR/MetaMap and SemRep production and development directories. Also, note that data directories should point to the directories referenced in SAWenv. The rule of thumb is that, in case of discrepancy between values in two files, SAWenv probably contains the correct setting.

2. *linkSICStus:* Used to compile SKR and it makes use of *SKRenv*. Settings regarding the *SKRenv* version to be used may need to be changed.

3. *s:* This script is used to debug SKR using the Emacs debugger. It makes use of *SKRen*v and the settings regarding the *SKRenv* version may need to be updated.

4. *make_all:* Used to compile C files. It makes use of *SKRen*v and the settings regarding the *SKRenv* version may need to be updated.

***prolog* directory:**

[1] This is draft. In the next version update, it should be carefully edited.

[2] There are a number of files in this directory prefixed as SKRenv, the version we are using as of SemRep v1.5 is SKRenv.12.

In this directory, two files may need to be changed (we may ask Francois to add these changes to SKR source code, since they really do not interfere with SKR/MetaMap).

1. *startup.pl*: maps module names to actual directories. Here, we need to ensure that *usemrep*, usemrep_* and *abgene* modules point to the correct directories. (*define_path* predicate). Also the following line may be added to determine_application/2 predicate, although I am not sure it really makes any difference:

   ; sub_atom('SAW', PWD) -> Area = usemrep

2. *init.pl:* The following lines were changed:

   environ('HOME', HOME), atom_concat(HOME, 'specialist/SKR/prolog/SICStus', HomePrologUtilsDir),

   to

   environ('SKR_HOME', HOME), atom_concat(HOME, '/prolog/', HomePrologUtilsDir),

   to make it work with current directory structure.

***src* directory:**

Main SKR source code resides in this directory. The changes in this directory are as follows:

1. *WSD/WSD* directory:

   1. *wsdmod.pl:* The changes in this file are meant for two purposes: first, exploiting preferred name/synonym exact match as WSD heuristics, and secondly, making Tokens information available for subsequent steps (positional information, most importantly). What original SKR returned for SemRep did not include this information. Since the changes in this file are plenty and scattered, we need to be careful while modifying it.

      1. export *do_WSD/8*, instead of *do_WSD/7*
      2. export *extract_SemRep_phrases_1/3*
      3. import built-in *lists:rev/2*
      4. import *metamap(metamap_tokenization:get_utterance_token_list/4)*
      5. import *skr_lib(nls_lists:get_from_list/3)*
      6. import *skr(skr:get_inputmatch_atoms_from_phrase/3)* and *skr(skr:get_phrase_tokens/4)*
      7. add argument Tokens to *do_WSD/8*
      8. add new predicate *preferred_name_synonym_equality* and the predicates it calls

      9. call *preferred_name_synonym_equality* from *do_WSD/8*

      10. call *get_utterance_token_list* from *do_WSD/8 before extract_SemRep_phrases/3, it returns TokensThisUtterance which becomes an argument for extract_SemRep_phrases/3.*

      11. add argument Tokens to *extract_SemRep_phrases/3* signature. The predicate itself has also changed.

      12. *extract_SemRep_phrases_1/3* definition added. There are several other new predicated called from *extract_SemRep_phrases_1/3*, as well.

2. *db* directory:

   1. *db_access.pl*: Since SKR uses a newer version of UMLS than SemRep does (2006), we must point to correct data files.

      1. modify to *default_version('SemRep').* Not anymore. Since DB.normal.06.strict also exists.

      2. *modify to default_year('06').*

      3. *modify to default_full_year(2006).*

      4. *Add FourDigitRelease == 2006 -> DefaultVersion='USAbase' to default_version.*

      5. *Replace default_release('2012AA') with default_release('2006AA').*

3. *lexicon* directory:

   1. *functions/c_linfl.c:* The change in this file is to fix a lexical access bug. The fix should be in future SKR releases, as well. I note it here anyway for reference.

   2. *Include/lm.h:* The same as above.

   3. *lexicon/qp_lexicon.pl: use_single_word_lexicon/0* predicate defined and exported.

4. *lib* directory:

   1. *semtype_translation06.pl:* This file does not exist in SKR, since it is no longer used, but we need it. It should be copied from SemSKR.

   2. *semnet_access06*.pl as well.

   3. *efficiency.pl:* export *maybe_atom_gc/3*, used by SemRep.[3]

   4. *nls_io.pl:*

[3] Could potentially be replaced with *maybe_atom_gc/2*

    1. define and export *fget_lines_until_null_line/2*. Used by SemRep and removed from SKR code since it is no longer used by SKR. [4] Replaced the SemRep call with fget_lines_until_skr_break/2 instead.

    2. comment out export for *fget_non_null_line/2.*[5]

  5. *nls_strings.pl:* define and export *trim_all_blanks/2*.

        Changed with trim_all_whitespace/2.

        Removed trim_all_whitespace/2 to ssuppserv.pl.

  6. *nls_lists.pl:* export *get_from_list_nd/3.*[6]

  7. *nls_system.pl:* The changes in this file involve adding new control options and removing obsolete ones.

    1. remove *filter_mrconso:m, mmi:M*

    2. change definition of *usemrep:D* from *dimitar_format* to *dysonym processing.*[7]

    3. Added usemrep:F, full_fielded_output, usemrep:R (write_syntax)

    4. remove *usemrep:m_mmofile*

    5. remove *usemrep:p (prolog_format)*

    6. remove *usemrep:Q (add_MMO_filename)*

    7. remove *usemrep:R, usemrep:W (read_smo_data, write_smo_data)*

    8. remove *usemrep:Y, usemrep:Z (read_smo_file, write_smo_file)*

    9. remove *metamap:B (moderate_model)*

    10. remove *skr_pvm_3map:M (moderate_model)*

5. *nls_text.pl:* Added hyphen as a graphical character. Is_graphic(0'-). [8]

[4] Could ask Francois to reinstate it, since it does not interfere with SKR.

[5] Could be removed.

[6] This was in red, but I think it is still valid. semgeninterp and semspec use this. We should eventually get rid of semgeninterp, but semspec will stay.

[7] Overall, check out all SemRep options.

6. *metamap* directory:

    1. *metamap_tokenization.pl:* define and export *get_utterance_token_list/4.* change in add_tokens_to_phrase_item. (get_subitems_feature)

    2. *metamap_utilities.*pl: change UMLS version.

        1. Modify to *skr_umls_info06*

        2. Modify to *semtype_translation06.*

7. *mmi* directory:

    1. *mmi.pl:* modify to *skr_umls_info06. It seems that this was wrong all along. It was really semtype_translation06, that was replaced, but this is no longer called anyway, rendering the whole change in this file unnecessary.*

8. *skr* directory:

    1. *skr_umls_info06.pl:* This file is no longer in SKR, so should be copied.[9]

    2. *skr.pl:* The changes in this file mainly relate to dysonym processing. In addition, several high level changes are made.

        1. *exclude_dysonyms* (entry point for dysonym processing) is defined and called.

        2. *skr_umls_info06* is now called.

        3. *skr_phrases_internal/4 is* defined and exported. This used to exist as metamap_internal in previous versions of SKR, but since it is not used by SKR, it was removed. We reinstated it as skr_phrases_internal. [10],[11]

        4. export *get_phrase_tokens/4 and get_inputmatch_atoms_from_phrases/3. Note that get_inputmatch_atoms_from_phrases/3 becomes get_inputmatch_atom_from_phrase/2.*

        5. *do_WSD/7* references are changed to *do_WSD* /8 and RawTokensIn is added as an argument. It is do_WSD/11 now, with WSDServerHost, WSDForced, WSDServerPort arguments.

[8] I think this was done to handle some metamap tokenization issues (well-established, etc.), but may not be necessary anymore. (Did not do this).

[9] Not sure versioned modules of skr_umls_info are used anymore.

[10] We can ask Francois to reinstate it for SKR in general, as well.

[11] Seems like this is replaced with skr_phrases/18 now?

        6. Import *usemrep_main:semgroup_member(semrep_semgroup_member/2)* predicate.
        7. Import *metamap:metamap_tokenization(get_utterance_token_list/4)* predicate.
        8. Import *nth0/3 (lists), midstring/6 (from sicstus_utils.pl)*
        9. Change the location of *import metamap:metamap_evaluation(matching_token).*
        10. Define *strings_to_atoms/2, atom_to_list/3, extract_mappings_from_maps/2*. *(from ssuppserv).*
    3. *skr_fe.pl:*
        1. export *form_original_sentences/7 and form_expanded_sentences/3.*
        2. Import *skr:nls_strings(trim_all_blanks/2).* [12]
    4. *skr_utilities.pl:*
        1. modify to *skr_umls_info06.*[13]
        2. Add format statement for spacing in machine output. (*write_MMO_terms_aux/1*).

9. *text* directory:
    1. *text_objects.pl:* Bug fixes. Removed *break_punc* and *hyphen_punc* for sentence boundary fix. Also, to fix an acronym problem, *match_initial_to_char* was change regarding *prep_conj_det*, disallowing them as acronym elements. [14]

[12] There is trim_whitespace/2 from nls_strings. I think the same exists in ssuppserv.pl. Need to consolidate.

[13] Only skr_umls_info exists.

[14] This was based on communication with Francois, and it might be already in new versions of SKR. Check that this works with “acute myocardial infarction” example.