Hi Brain:

Thank you very much for sending us the testing results of lvg (LuiNorm). As usual, we analyze these three result files to monitor luiNorm's behavior and enhance it (if needed) for the next release. This year, as expected, there are more merge (7482) than split (183) and split_merge (93). This is because of 1) enhanced canonical algorithm converts spelling variants into same canonical classes; 2) more data on spelling variants in LEXICON; 3) More words coverage (in atoms.data and LEXICON). The main cause (of all three files) is the difference of canonical classes between 2006 and 2005. In general, a canonical class in 2006 covers bigger range (more words). For examples, spelling variants with '-' and ' ' result in the same canonical class in this release. I went through data on all three files and found luiNorm behave the way it should be. Bellows are the detail explanations if you are interested:

==========

1) Merge: 15068 Luis merge into 7482 Luis.

1-a). Merged because of spelling variants with '-'

This case includes 53.53% (4005/7482) of all merge cases. Reasons of this merge are:
- Enhanced canonical algorithm converts all spelling variants into same canonical classes
- New spelling variants with '-' are inserted in the LEXICON.
- New words with '-'

For example,
L0000032|L0000032|S0102903|1,2-benzopyrone
"Benzopyrone" has spelling variant of "benzo-pyrone" and belongs to the same canonical class in 2006. Thus, after sorting the canonical forms by ASCII order in luiNorm, "Benzopyrone", "benzo-pyrone", and "benzo pyrone" return the same luiNorm result of "benzo pyrone".

1-b). Merged because of spelling variants with ' '

This case includes 33.48% (2505/7482) of all merge cases. Reasons of this merge are:
- Enhanced canonical algorithm converts spelling variants into same canonical classes
- New spelling variants with ' ' are inserted in the LEXICON.
- New words with ' '

For example,
L0001758|L0001758|S0325164|aftercare
"aftercare" has spelling variants of "after care" and "after-care" and all of them convert to the same canonical form. Thus, they all have the same luiNorm result of "after care".

1-c). Merged because of new spelling variants, new words (inflections), or spelling variants with other punctuation:

This case includes 12.99% (972/7482) of all merge cases. Mainly, there are three cases:
- new spelling variants (new records) in 2006 LEXICON Such as in L5854441, "minocyclin" and "minocycline" are new spelling variants in 2006 LEXICON. Thus, they have same luiNorm result of "minocyclin".

- new words (inflection) in 2006 LEXICON:
Such as in L0000919, the variants=reg is added into "proneness" in 2006. Accordingly, "pronenesses" belong to same canonical form. Thus, they have same luiNorm result of "proneness".

- spelling variants with punctuation:
Such as in L5862327, "CB" has new spelling variants "C.B." in 2006. Accordingly, "CB" and "B with C" have same luiNorm result of "b c".

==========
2) Split: 182 Luis split into 368 Luis

2-a). New lexical records:

This case includes 59.56 % (109/183) of all split cases. New lexical records might result in different inflectional variants and convert to different canonical form. Such as L0002153, "alloxanthine" is in LEXICON, 2006, with "variants=uncount". The inflectional variants are generated based on the FACTs (2006) instead of RULES (as in 2005). "alloxanthine" and "alloxanthin" belong to two different canonical classes in 2006 because they don't share any same inflectional variables (alloxanthined) as in 2005. Thus, they have different luiNorm result.

2-b). Change of lexical records and atoms.data:

The rest of 40.44% (74/183) belongs to this case. Lexical records are changed (in inflectional variants and spelling variants) and results in different inflectional variants to convert into different canonical forms. For example, "vitamine" is not a spelling variant of "vitamin" in LEXICON 2006. Accordingly, L55866482|Vitamine splits from L0042890|Vitamin.

The change (deleted terms) in atoms.data also could cause the split. For example, "kappa Alltype b4" and "kappa Alltype b4s" are deleted from atoms.data, 2006. Thus, there is no canonical class includes "b4s" in 2006. This caused the split.

There are some cases in this category that I think it should not split, but it did by the current algorithm. For example, "Cygnoline", "Cygnolin", and "Cignolin". These records should be converted into one canonical form because they share same word "cygnoline". However, they have different luiNorm results of "cygnolin" and "cignolin" because they have different citation forms. I will look into it in the future release.

==========
3). Split-Merge: 93 terms in 24 split-merge cases

The cause of this case is the combination of above two (Split and merge). Potentially, terms in these 24 cases might belong to same canonical class. This is the data I will use to enhance the algorithm of canonical form and make canonical class covers bigger range (more words) when it makes sense.

3-a). New lexical records
* There are 37.5% (9/15) in this split_merge cases caused by new lexical records. For example, L0144613|pep is a new record in LEXICON with "variants=regd". This results in different inflectional variants than those in 2005 (generated by RULEs). This causes different luiNorm results and the split.

3-b). Others
* 15 out of 24 in split-merge cases are caused by stripping ambiguity tags, modification of lexical records (spelling variants, inflectional variants).

Also, I found that -f:rs flow does not remove (S), (ES), (IES) when it is upper case in luiNorm. There are few cases like this. Such as L5318868, L5302468, etc.. It seems to me to remove upper cases parenthesis plural forms make more sense. I will also look into it in the future release.

Hope this makes sense! Thank you!

-- Chris

==========

Chris,

Each time LVG versions change we run a sample LUI re-assignment to test the new algorithm (as discussed in recent meetings). The result of this is three reports showing the splits, merges, and split-merge cases resulting from the new LVG. Each report has the form OLD_LUI|NEW_LUI|SUI|STRING and it shows how LUIs from the previous version were changed in the current version.

The three reports are attached. There were 7482 merged LUIs, 183 Split LUIs, and about a hundred of the complex split-merge case.

Following are 1 example from each report:

split_merge.dat

Here L0056663 split partially and partially merged with L0803694:
L0056663|L0056663|S0146976|cyanomethemoglobin|
L0056663|L0056663|S0831228|CYANOMETHEMOGLOBIN|
L0056663|L0803694|S0611612|CYANMETHEMOGLOBIN|
L0056663|L0803694|S0852119|Cyanmethemoglobin|
L0803694|L0803694|S0831222|CYANMETHAEMOGLOBIN|
L0803694|L0803694|S0852117|Cyanmethaemoglobin|

merge.dat

Here L5163138 is merged into L0000032.
L0000032|L0000032|S0007557|1,2-Benzopyrones
L0000032|L0000032|S0102903|1,2-benzopyrone
L0000032|L0000032|S0463008|1,2 Benzopyrones
L5163138|L0000032|S5923205|1,2 Benzo Pyrones
L5163138|L0000032|S5923265|1,2-Benzo-Pyrones

split.dat

Here L0001626 split into L0001626 and L5866329
L0001626|L0001626|S0011240|Adrenal Glomerulosa
L0001626|L0001626|S0045050|Glomerulosa, Adrenal
L0001626|L5866329|S0011241|Adrenal Glomerulosas
L0001626|L5866329|S0045052|Glomerulosas, Adrenal

Brian