Language Learning Diary - Part Two

Linas Vepstas

2021-present

Abstract

The language-learning effort involves research and software development to implement the ideas concerning unsupervised learning of grammar, syntax and semantics from corpora. This document contains supplementary notes and a loosely-organized semi-chronological diary of results. The notes here might not always makes sense; they are a short-hand for my own benefit, rather than aimed at you, dear reader!

Introduction

Part two of the diary on the language-learning effort starts with the new task of of closed loop learning. The idea of closed loop learning is that the accuracy of the learned grammars can be very closely monitored and measured, thus allowing the learning algorithms to be tuned for speed and measured for accuracy. The closed loop is a basic three-step process:

1. Generating random but controlled grammars
2. Generate a text corpus from these grammars
3. Learn a new grammar from the corpus
4. Compare the controlled-grammar to the learned-grammar
5. Tune algorithms and procedures, and repeat.

The step-by-step instructions can be found in the file "README-Calibration.md".

February 2021

Restart the project, finally making some headway. ToDo items:

- Describe how the "uniform random sampling of sentences" is performed.
- Enable weighted random sampling of sentences.
- Fix multi-sense bug in gen-dict.scm circa line 85

First run. Instructions in "README-Calibration.md". Lots and lots of bugs fixed and lots of pipeline was automated.

March 2021

First real experiment, in expt-7/expt-8/expt-9. The dict in expt-7 fails to generate the correct corpus, because the generator does not expand synonyms (this is a combinatorial explosion and it just doesn't do that.) Expanded by hand in expt-8. Same data, new scripts and config in expt-9.

Issues:

- After MST parsing, the grammar is correct, in that (I think) it will produce exactly the same corpus. However, the rules are different (and more verbose). TODO: check that the same corpus is actually produced. How? Answer: generate the corpus, and compare...
- After MST (MPG) parsing, the verbs link to obj-determiner instead of object. Why? Was there a cutoff that was missed? The corpus is just .. tiny.
- After (gram-classify-greedy-discrim 0.5 4) the right clusters are produced, but the connectors are not clustered; they need to be. TBD.
- Result: the correct corpus is produced (via manual checking), however, extra sentences are produced, which are missing the verb. This is because the wall links to the two determiners, and the wall can be skipped. However, the sentence with the verb has a higher MI. (11.51 instead of 8.92. See below.) This is due to a bug. See below.

Here's the result:

```
linkparser> the squirrel a dog
Found 1 linkage (1 had no P.P. violations)
        Unique linkage, cost vector = (UNUSED=0 DIS=-8.92 LEN=2)

     +-----------TB-----------+
     +---TB---+---TE---+      +-TE-+
     |        |        |      |    |
LEFT-WALL.2 the.1 squirrel.3 a.1 dog.3

linkparser> the squirrel saw a dog
Found 1 linkage (1 had no P.P. violations)
        Unique linkage, cost vector = (UNUSED=0 DIS=-11.51 LEN=4)

     +--------------TB--------------+
     |                 +-----TC-----+
     +---TB---+---TE---+---TF--+-TD-+-TE-+
     |        |        |       |    |    |
LEFT-WALL.2 the.1 squirrel.3 saw.4 a.1 dog.3
```

Hypothesis: The MI of wall-verb is lower than the MI of wall-determiner. Thus, the planar parser always picks the wall-determiner. Lets find out.

Heh. There is no MI wall-verb! Ouch. This was due to bad sampling; fixed github.com/opencog/opencog commit 895226228 Mar 16 2021. Sheesh.

Expt-10

expt-10, this is fixed. The parse trees are now very rich. Generated sentences:

- length 4 - none
- length 5 - the expected ones.
- 6 - LEFT-WALL LEFT-WALL plus valid sentence
- 7 - none
- 8 - LEFT-WALL the mouse saw the dog chased a bird
- 9 - double left wall
- 10 - the LEFT-WALL a cat saw the dog chased a squirrel – and also a triple-left-wall.

WTF. what's with the crazy multi-left-wall!? Heh. Here we go:

```
     +---------------------TI-------------------+
     +------------------TB-----------------+    |
     |           +------TI-----+           |    |
     |           +---TB--+--TC-+--TO--+-TE-+-TC-+
     |           |       |     |      |    |    |
LEFT-WALL.2 LEFT-WALL.2 a.1 mouse.5 saw.3 a.1 cat.5
```

Expt-11

So... expt-11 places a period at the end of every sentence. That terminates the infinite-recursive lengths being generated to only finite-length sentences. There is a total of three different parses. All have exactly the same cost. These are as follows:

```
Found 3 linkages (3 had no P.P. violations)
Linkage 1, cost vector = (UNUSED=0 DIS=-15.67 LEN=10)

     +-------------------TH-------------------+
     |              +------------TJ-----------+
     +------TG------+              +----TE----+
     +---TF---+--TD-+---TI--+--TC--+--TD-+-TJ-+
     |        |     |       |      |     |    |
LEFT-WALL.2 the.1 dog.3 chased.4 the.1 cat.3 ..5
```

```
      Linkage 2, cost vector = (UNUSED=0 DIS=-15.67 LEN=11)
    +-------------------TH-------------------+
    |               +------------TJ-----------+
    +------TG------+------TB------+----TE----+
    +---TF---+--TD-+---TI--+--TC--+--TD-+-TJ-+
    |        |     |       |      |     |    |
LEFT-WALL.2 the.1 dog.3 chased.4 the.1 cat.3 ..5

         Linkage 3, cost vector = (UNUSED=0 DIS=-15.67 LEN=12)

    +-------------------TH-------------------+
    |               +------------TJ-----------+
    |               |       +--------TK-------+
    +------TG------+       |      +----TE----+
    +---TF---+--TD-+---TI--+--TC--+--TD-+-TJ-+
    |        |     |       |      |     |    |
LEFT-WALL.2 the.1 dog.3 chased.4 the.1 cat.3 ..5
```

Notable in the above:

- Only one parse, the third one, links the main verb to a wall. And then its the right wall, not the left wall.
- The determiners seem to be over-linked, and judged to play a too-important role.
- The output grammar is much more highly detailed and constrained than the intended grammar.

Questions:

- What happens if there are a lot more verbs? Would this make the determiners less important, more important, or have no effect? (My guess is “no effect”)
- To downgrade the importance of determiners would seem to require having sentences without them in it.

TBD:

- Waiting on completion of link-generator so that multiple-sense corpora can be generated. (enabled in lg pull req #1175) Or something like that ... what is the right strategy here? Need to rethink to avoid combinatorial explosion, while also verifying category contents.
- Fix bug to allow multiple-sense word definitions in multiple dict locations.
- Dict generation should auto-handle placing a period at the end of the sentence.

Expt-12

Start work on a single-sense random dictionary. Hit assorted issues with the scripts.

Results:

- non-classified dict has 134391 disjuncts, 11 uni-classes including left-wall. These are raw disjuncts.
- classified dict (i.e. that on which grammatical classification has been performed) has 11286 disjuncts

Time to generate 50 sentences, and all possible sentences, in seconds.

length	time for 50	time for all	num sents
3	3	3	108
4	3	3	779
5	4	4	7107
6	5	8	67935
7	13	31	673812
8	43	370	6855920
9	140		
10	638		
11	963		
12	2180		
13	3364		
14	5850		

Data is semi-meaningless, scripts were broken, data processing would start before data was fully loaded. Try again. Upon restart, the number of sentences generated is order of magnitude lower. Presumably due to corrected clustering; above clustered incomplete lists of disjuncts and thus over-generalized.

Expt-13, expt-14, expt-15

Try again with the same initial corpus. Expt-13 overflowed with fake warning message, so I couldn't see the log; thus expt-14 is an exact rerun. "Exact" in the sense of using the same config. However, the random sampling of pairs during pair counting was different.

Then expt-15 uses exactly the same corpus, with a period at the end of sentence placed manually.

Columns:

- length: length of sentence
- time to generate all sentences, in seconds (expt-13)
- num: number of sentences generated (expt-13)

- expt-14: number of sentences generated
- corpus: number of sentences in input corpus. Capped at 25K sentences for the longer sentences. Second number is how many could have been generated.
- expt-15: redo, but with a period at the end of the sentence.

length	time for all	num expt-13	expt-14	corpus	expt-15
3	3	19	21	10	24
4	2	142	163	23	104
5	2	1130	1356	75	485
6	2	9732	12090	254	2294
7	5	86872	111633	892	10845
8	13	794320	1054583	3402	51673
9	143	7393748	10134151	12728	248242
10	2558	69781807	98702133	25000/48364	1198418
11				25000/187541	5807783
12				25000/733525	28246686

Expt-13 and expt-14 differ only in how the pair-counts were collected (they are randomly different samplings of pairs). Both wildly over-sample the corpus.

The expt-13/14 input corpus lacks periods at the ends of sentences. This seems to be the most likely explanation for the over-generation; i.e. last time a period was lacking, the same thing happened. So expt-15 takes exactly the same corpus - identical copy, and adds a period. This does sharply cut down on the number of sentences, especially the long ones, but still over-generates.

General processing stats:

	expt-14	expt-15
time, pair-counting	109 minutes	107 minutes
pair aid	236/293	259/318
pair dimensions	11 x 10	11x10
pair counts	27367456	29988408
pair sparsity	0.0	0.0
pair entropy	5.96=2.90+3.06	6.14=3.08+3.08
pair MI	0.0014	0.012
time, mpg-parsing	12 minutes	19 minutes
mpg aid	246691	201594/298813
mpg dim	11 x 111712	12 x 97203
mpg counts	1067417	1159801
mpg sparsity	3.19	3.49
mpg MM^T support	134663	104049
mpg MM^T count	2341493237	9220550155
mpg MM^T entropy	2.95	0.232
gram dim	3 x 103552	4 x 92299
gram count	880090	1080580
gram sparsity	1.60	2.46
gram entropy	11.58=11.30+1.09-MI	8.83=8.75+1.15-MI
gram MI	0.806	1.08
dict records	102644	67140

Issues:

- There seems to be a dataset issue: both the disjunct pairs and gram pairs have 1/3rd of them without counts on them. .. They are not being saved, after clustering (clustering causes deletion of many disjuncts, and alteration of counts on all disjuncts.) I'm guessing this failure results in bad MI's? Anyway, its a bug is fixed in the new clustering shell scripts.

After above fix, verify export. gram-1 is a re-export of the original expt-13 run, while gram-4 is export of the fixed run. (Both start with the same disjuncts. I think there's nothing stochastic/random during processing, so it should be repeatable...)

	expt-13/gram-1	expt-13/gram-4
gram dim	3 x 103325	4 x 111320
gram count	881340	939886
gram sparsity	1.60	1.92
gram entropy	11.57=11.29+1.09-MI	11.89=11.61+1.17-MI
gram MI	0.806	0.894
dict records	101922	117807

So .. similar but not the same. How about sentence generation? expt-13-gram-1 and expt-13-gram-1a are identical (the 1a version is from a re-export; so we're exporting the same stuff).

The classes in gram-1 are <b f> <e j> and left-wall. The classes in gram-4 are the same plus <i#uni> ... there was no word i in gram-1 !! Wow, that's a big drop.

length	time for all	expt-13-gram-1	expt-13-gram-4
3	3	19	19
4	2	142	163
5	2	1130	1414
6	2	9732	13147
7	5	86872	125527
8	13	794320	1225346
9	143	7393748	12156101
10	2558	69781807	122141737

OK, so the numbers are dramatically larger. Apparently, this is due to the previously dropped word i. Yikes!

Well, the above is massively under-counting – it is only sampling one random word-draw per class. Since multiple words are in each class...

... anyway, this is nuts, because the connectors need to be classified, instead of issuing new connector types. So more work before something meaningful is possible.

expt-13 vs expt-15 precision, recall

So now that we've got things working, lets look at precision and recall. Clearly precision will be terrible, but maybe recall will be excellent? Compare expt-13-gram-4 to expt-15-gram-2 (which rebuilds after fixing the borked save.

Uhh .. No its not working yet, in the sense that the connector classes are being mis-handled. Those need to be grouped correctly before export. More work...

expt-16

Due to absence of left-wall in the above, breaking the dict-compare step, tried again, generating a new dict with walls (after manually adding a wall to the dict of expt-15, and ending punctuation to the dictionary.) Well ... clustering worked quite differently. Here's a summary.

Pair counting seems to be more-or-less the same, slightly higher MI=0.030 which is still minuscule. MPG entropy, counts etc. look similar to the earlier runs.

Gram classification: only 2 words assigned to the same class. Oh, this used the "disinfo" classifier, whereas the earlier runs used the "fuzz" classifier. That could account for everything, I guess. Lets take a look.

	disinfo 3.0 4	discrim 0.5 4	fuzz 0.65 0.3 4
gram dim	11 x 75552	7 x 75667	7 x 75667
gram count	1011444	951215	960646
gram sparsity	3.37	2.77	2.77
gram entropy	11.48=10.01+3.15-MI	10.38=9.94+1.90-MI	10.45=9.91+2.06-MI
gram MI	1.68	1.46	1.52
dict records	80460	77735	77750

Clearly, the resulting clusters are sensitive to the parameters controlling classification. The above parameters seemed reasonable for the large English dataset. They may be unreasonable here!? But this is very unclear.

The MI's are larger, across the board (vs. 0.8 or 0.9 before.) How about sentence generation? We expect disinfo to be more accurate, since it did very little clustering.

length	time disinfo	disinfo	time discrim	discrim
3	61	60	3	18
4	67	523	3	107
5	131	5217	3	752
6	137	51368	3	5371
7	183	488376	5	37923
8	715	4514440	8	268248
9			37	1909447
10			262	13616410

Again, the discrim is under-counting, because of more categorization.

Next step: fix the conjoined clustering, with shapes.

Wow. So MM^T entropy with shapes is 5.03 which is huge compared to the dj-only MM^T so it really is something new and different! With shapes, and with gram-disinfo, there were no merges.

	disjunct disinfo 3.0 4	shape disinfo 3.0 4
MM^T MI		5.03
gram dim	11 x 75552	12 x 75552
gram count	1011444	1015356
gram sparsity	3.37	
gram entropy	11.48=10.01+3.15-MI	11.50=10.02+3.18-MI
gram MI	1.68	1.69
dict records	80460	80807

OK, so looks like shape created no categories at all. So how does that work out for generation?

```
link-generator -l learned -c 123123123 -s 3
```

length	corpus	time for all	expt-16-shape
3	4	38	61
4	21	69	566
5	50	147	5638
6	179	124	55546
7	621	172	531075
8	2246	642	4937036
9	8850		
10	> 25K		

OK, so wildly over-generating sentences, despite effectively no clustering being done. Didn't we do an experiment without clustering?? I can't find it above. Why are we over-generating? How to best explain it? Too small a vocabulary?

- Issue 1: given a fake-lang the generator is failing to generate all possible sentences. Fixed in link-grammar pull req #1175
- Issue 2: there are "accidental" synonyms cause of 1 above: many POS'es are shared in common between many words but are not completely sampled, thus creating "accidental synonyms".

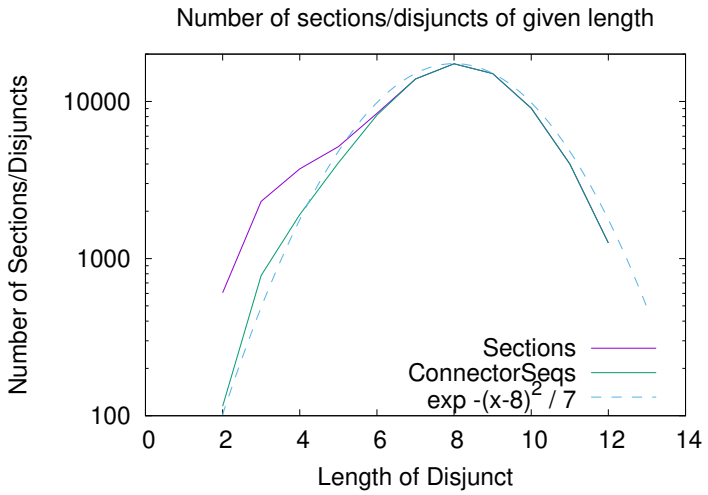
OK, so back to that one, except this time, its in the shape of a bug...

Sections and disjuncts

The Sections that were learned in expt-16 have a surprising number of connectors on them, averaging at 7.7 connectors per section. This seems way too large. What's up with that? Step one: get a more detailed view.

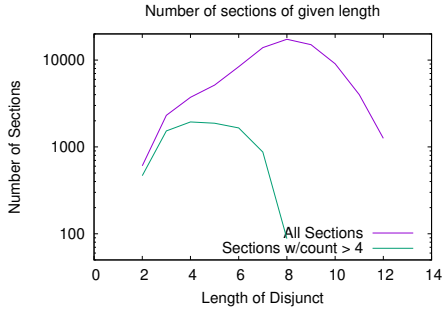
```
# column 1 length
# column 2: number of Section
# column 3: number of ConnectorSeq
#
1  0  0
2  606   115
3  2311  780
4  3723  1904
5  5150  4066
6  8387  8183
7  13907 13896
8  17362 17362
9  15043 15043
10 9061  9061
11 4000  4000
12 1257  1257
13 0
```

So ... long disjuncts appear on one and only one word (as witnessed by identical counts for length 8 and above). Short disjuncts might be shared with multiple words. This implies that words are not mergeable, or rather, the distinct long sections are preventing merger. Here a graph of above.



So its very Gaussian, both peak and tails. All these gaussians here and in earlier results imply that central-limit theorems hold. And all kinds of other classical theorems should hold. Have not been leveraging those theorems, so far. If we did, what would we get? Note it's Gaussian despite the fact that negative connector-seq lengths don't make sense! Right? What would a negative connector-seq length be? A repulsion? an anti-connect statement? "Must never connect"?

What happens if we exclude Sections with a low observation count?



The above shows all sections, and sections that were observed 5 or more times. Apparently, the long, complicated disjuncts are observed very rarely.

How should this be interpreted? Common sense seems to suggest that low-count observations are "noise" and should be cut before any merging is performed. Doing so will certainly increase the similarity of vectors. But are cuts really needed? Perhaps

the similarity measures can already deal with these? If so, then the only reason for cuts would be performance, rather than accuracy.

How do these affect cosine-similarity vs MI-similarity? Clearly, for cosine-similarity, low counts means short vector components, and so these will not contribute much to the dot product. Likewise, the MI-similarity is built on a dot-product, so again, these should not contribute much. Even more-so, since the MI-similarity never takes a square-root of the dot product. It does, however, re-weight basis elements in a fashion that I do not yet have a good intuition for. (Or rather, developed an intuition, and now I've forgotten what it was... Hmmm.)

TODO: The Shapiro–Wilk test can be used to determine how close a distribution is to a Gaussian. The Wikipedia article on it is as clear as mud.

April-May 2021

Before starting expt-17 with fixed sampling, it is time to ponder connector merging. This is turning out to be non-trivial. The general upshot of the below is that the original 2019 concept for Shapes and CrossSections is a good, strong idea. However, assorted non-commutativities arise, and how to resolve them appropriately is not entirely clear. Thus, we embark on a journey of discovery...

But first: connector merging is needed because without it, the clustering fails to adequately reduce the size of dictionaries. The meta question is this: if words A and B are determined to be near-synonyms, and assigned to the same cluster, then what should be done with connectors that have A or B in them? Should all such connectors be automatically replaced by the cluster? The naive answer is, "yes, they should be".

The less-naive response is "but multiple word-senses". That is, A and B are not just words, but are word-vectors, and vector B may be a linear combintation of two different word senses. One of these senses might be synonymous to A, and the other might be completely diffferent. Thus, we want to merge that part of vector B that is (nearly) colinear with A, while leaving behind a different vector B-prime that is associated with some other word-sense for the word B. When encountering B in a connector, is that B in the sense of wordclass-AB, or is it in the sense of B-prime? If the former, then obviously, that connector should be updated to use wordclass-AB; otherwise, it needs to be left alone, thus implicitly defaulting to B-prime.

TBD: Writing the above, it appears that the need for an explicit B-prime has been overlooked. This could cause havoc, in allowing inappropriate linkages! This is important, and needs to be addressed ...

Connector merging

Some of the questions that arise with connector merging, to figure out:

- How would connector merging affect clustering results?
- If connectors are merged, then how should vectors be handled? (Naive vectors no longer work, because the basis is now different.)

- What is the best or most correct merge algorithm?

Lets look at a toy model. Suppose the dict is

```
A: C+ & D+
B: C+ & D+

C: A- & E+
C: B- & F+

D: A- & G+
D: B- & G+
```

From this, conclude that A and B can be merged. However, the connectors on C cannot be merged, the connectors on D can be merged. The final dict is then (with some abuse of notation)

```
<wclass-AB>: C+ & D+
A B: <wclass-AB>

C: A- & E+
C: B- & F+

D: <wclass-AB>- & G+
```

Next suppose that we have

```
S: C+ & D+
T: C+ & D+

C: S- & J+
C: T- & F+

D: S- & H+
D: T- & G-
```

The determination to merge is now trickier. Naively, S can be merged into `<wclass-AB>` but doing so would wreck the connector set on D, as it implies an `(S- & G+)` which was not observed, and thus not mergeable. ... unless we are willing to create such new unobserved cases.

Merging T looks OK.

Perhaps this looks bad because we are not doing shape vectors. The shape vector for the above would be:

```
A: (C+ & D+) or (C:x- & E+) or (D:x- & G+)
B: (C+ & D+) or (C:x- & F+) or (D:x- & G+)
```

```
C: (A- & E+) or ...
C: (B- & F+) or ...

D: (A- & G+) or (A:C+ & x+)
D: (B- & G+) or (B:C+ & x+)
```

So the decision to merge A and B is less clear-cut, when using shapes (they are not perfect synonyms). The shape variant of the S thing is

```
S: (C+ & D+) or (C:x- & J+) or (D:x- & H+)
T: (C+ & D+) or (C:x- & F+) or (D:x- & G+)

D: (S- & H+) or (S:C+ & x+)
D: (T- & H+) or (T:C+ & x+)
```

So using shapes weakens the decision to merge S into AB. The decision to merge T remains strong. So it seems that shapes do offer a stronger foundation on which to make merge decisions. They examine similarity (almost-synonymy) out to a greater distance from the germ. It seems like they also allow things to continue to be treated as vectors, instead of muddling the concept of vectors. After the AB-merge[1] the result would be:

```
<wclass-AB>: (C+ & D+) or (C:x- & E+) or (C:x- & F+) or (D:x- & G+)
A B: <wclass-AB>

C: A- & E+
C: B- & F+

D: (<wclass-AB>- & G+) or (<wclass-AB>:C+ & x+)
```

Now, given this merged-AB thing, when happens when we look at merging S? Well, S is meh, T looks better. The vector for T is comparable to the vector for <wclass-AB> so vector similarity works for that merge decision.

Conclude: the original plan from a few years ago works and holds water. Use shape vectors for merge decisions. Once this is done, connectors can be swept up.

Non-commutativity

The above description pulls a sleight-of-hand, which presumes an algorithm that is able to crawl across individual disjuncts, compare them, and update words with merged word-classes. Such an algorithm can be written (and has been written/prototyped). It leads to some confusion, because the shapes/cross-sections are no longer consistent. Lets call the above the "connector sweep algorithm", or "sweep" for short.

[1] When using the "union-merge" strategy, as described in in `src/gram-projective.scm`. In practice, the merge style used is typically `merge-project`, which would accept only a fraction of `(C:x- & E+)` and `(C:x- & F+)` into the final vector.

Starting with the above example:

```
<wclass-AB>: (C+ & D+) or (C:x- & E+) or (C:x- & F+) or (D:x- & G+)
```

the sections can be reconstructed from the cross-sections. The reconstructed sections are

```
C: <wclass-AB>- & E+
C: <wclass-AB>- & F+

D: <wclass-AB>- & G+
```

Comparing, the reconstructed section on D is the same as what the sweep algo produced, but the sections on C are *not* what the sweep merge offers. That is, the sweep is not commutative with the creation of shapes. This is a problem for maintaining the consistency between sections and cross-sections as clusters grow. The lack of consistency will cause merge judgements to diverge...

Thus, we have at least two algorithms:

- Sweep-merge, as described above, where connectors are replaced by merged-connectors if and only if the the rest of the connector set is identical. This merge algorithm is naively described, since it does not explain what to do if there are multiple connectors in a conector set that might be merged. It's also naive in that it does not explain how counts (frequencies) are to be handled.
- Reshape-merge, which performs the basic projective merge on the germ-vectors, and then reconstructs Sections from CrossSections, thus restoring consistency between sections and cross-sections. It violates the intuitive correctness of the sweep-merge, but only perhaps because the sweep-merge, as naively described above, assumed the "union-merge" strategy of transfering observation counts for vectors taht are not perfectly colinear. The projective-merge count transfers recognize the non-colinearity, and obtain cluster centroids through weighting formulas.

The above ruminations suggest that reshape-merge enjoys an advantage over sweep-merge, as it keeps the section/cross-section duality consistent.

Anyway, the original 2019 plan for using shapes seems to have been a good plan.

Non-Commutivity, Again

The non-communtivity can be hightened with a slightly richer example. Consider

```
A: (P+ & Q+) or (R+ & S+) or (K- & B+)
B: (P+ & Q+) or (R+ & S+)
C: B- & T+
```

The dictionary entry of C is present to remind us of the fact that, if B+ appears as a connector, then B- must also appear as a connector, somewhere.

Based on the first two sections, a decision might be made to merge, with the count on `(K- & B+)` being small enough that it does not disrupt the merge decision. Expanding this into it's CrossSections, the full vectors are:

```
A: (P+ & Q+) or (R+ & S+) or (K- & B+)
B: (P+ & Q+) or (R+ & S+) or (K- & A:x+) or (C:x- & T+)
```

The merge result is then

```
<wclass-AB>: (P+ & Q+) or (R+ & S+) or (K- & B+)
<wclass-AB>: (K- & A:x+) or (C:x- & T+)
```

The cross-section leads to a reconstruction (reshape) of

```
A: K- & <wclass-AB>+
C: <wclass-AB>- & T+
```

How is this to be interpreted? Let's explore some "common sense" reasoning.

Case A: The count on `(K+ & B+)` is so small that it is considered to be noise, and is completely dropped before merging even starts. In this case, A and B are exactly colinear (are exact synonyms). The (naive) merge of A and B is completly unproblematic, except that it leaves C without the ability to connect to anything. This can be handled in one of two ways. One way is to notice that, by detailed balance, the counts on this particular C section must also be tiny, and so this section can be dropped from the dictionary.

Another way to avoid this dangling-connector problem is to presume that the dictionary also includes

```
D: L- & B+
```

which would provide a place for that bar B to connect. but if this were the case, then we did the cross-sections on B wrong. Fixing these would have given

```
<wclass-AB>: (P+ & Q+) or (R+ & S+) or (K- & B+)
<wclass-AB>: (K- & A:x+) or (C:x- & T+) or (L- & D:x+)
```

which then reshapes to

```
A: K- & <wclass-AB>+
C: <wclass-AB>- & T+
D: L- & <wclass-AB>+
```

This is now fully linkable, and there are no dangling pure-B connectors.

In conclusion, this seems self-consistent eitehr way: either we can drop the `A:(K+ & B+)` section entirely, and, by detailed balance, we can drop `D:(L- & B+)` also; or we can keep both, and doing it correctly leaves nothing dangling.

Note that we have to be careful with tracking in the merge algo: when reshaping to get the `C:<wclass-AB>- & T+` section, we have to be careful to notice that the `C:B- & T+` section was a donor, and so it should be removed (its counts driven to zero). Otherwise, C would end with both these sections on it, and it would be a bit wonky.

Case B: The count on `(K+ & B+)` is small but not ignorable. It is small enough to not block the merge decision. There are two issues to resolve. The first is easy: the `C:B- & T+` section should be recognized as a donor to `C:<wclass-AB>- & T+`, and removed (its counts driven to zero). This is easy enough to determine at the time of the merge.

The more difficult issue is what to do about the `<wclass-AB>:(K- & B+)` section, which appears to have a dangling B connector, and the reshape of `A:(K- & <wclass-AB>+)`, which seems to be a double-count. The first leaves a dangling A, the second leaves a dangling B. It seems fairly clear that these should be harmonized, merged together, to give `<wclass-AB>:(K- & <wclass-AB>+)`. It seems that this can be reasonably infered and performed at the time of creation, since the donors are readily identified.

Case C: The count on `(K+ & B+)` is large, large enough to split. That is, A should be understood to be the direct sum of two distinct word-senses, with one word-sense being `<wclass-AB>` and the other being `<A-prime>:(K+ & B+)`. So, if were were able to be absolutely sure that A-prime was really a distinct word-sense, then we should transfer none of the counts from the originating section `A:(K+ & B+)` to the `<wclass-AB>` section.

So, starting with the vectors

```
A: (P+ & Q+) or (R+ & S+) or (K- & B+)[n]
B: (P+ & Q+) or (R+ & S+) or (K- & A:x+)[n] or (C:x- & T+)
```

where square-bracket-n is the count on that section, we create a merge result of the form

```
<wclass-AB>: (P+ & Q+) or (R+ & S+)
<wclass-AB>: (C:x- & T+)
<A-prime>:  (K- & B+)[n]
B: (K- & <A-prime>:x+)[n]
```

The cross-sections leads to a reconstruction (reshape) that appears to be self-consistent, so I don't see any problems here.

If we were to split [n] into some fractional parts, then this would reduce to a combination of case B and the current case, so that should also work.

Connector counts

(TBD, this section needs to be harmonized with the new text above... the below was written before the above was rewritten...) Lets go through above exercise, this time with counts. Suppose the dict, with observation counts in square brackets, is

```
A: C+ & D+ [na]
B: C+ & D+ [nb]

C: A- & E+ [nca]
C: B- & F+ [ncb]

D: A- & G+ [nda]
D: B- & G+ [ndb]
```

The shapes, with counts, are

```
A: (C+ & D+)[na] or (C:x- & E+)[nca] or (D:x- & G+)[nda]
B: (C+ & D+)[nb] or (C:x- & F+)[ncb] or (D:x- & G+)[ndb]

C: (A- & E+)[nca] or ...
C: (B- & F+)[ncb] or ...

D: (A- & G+)[nda] or (A:C+ & x+)[na]
D: (B- & G+)[ndb] or (B:C+ & x+)[nb]
```

Merging A and B, with 100% of count transfer, gives

```
A B: (C+ & D+)[na+nb] or (C:x- & E+)[nca] or (C:x- & F+)[ncb] or (D:x- & G+)[nda+ndb]

C: (A- & E+)[nca] or ...
C: (B- & F+)[ncb] or ...

D: (A- & G+)[nda] or (A:C+ & x+)[na]
D: (B- & G+)[ndb] or (B:C+ & x+)[nb]
```

Looking at the connectors on D, we see that they are mergable, and that the counts are consistent. So, based on this toy model, we can either try to merge connectors directly, or we can, at a later date, merge connectors by reconstructing them from merged shapes. Doing it either way should give the same counts: the operations are commutative.

This is not entirely obvious. It seems to work for the toy example. It seems like the toy example could be converted into a full proof. Yet ... are we missing something? Best bet is to write the code both ways, and very numericall that the operations are commutative.

Fractional counts

Lets try again, this time with fractional counts. Suppose that instead of merging 100% of B into A, we merge only a fraction $0 \leq y \leq 1$ of the count. This gives

```
A B: (C+ & D+)[na+y*nb] or (C:x- & E+)[nca] or (C:x- & F+)[y*ncb] or (D:x- & G+)[nda+y*ndb]
B: C+ & D+ [(1-y)nb] or (C:x- & F+)[(1-y)ncb] or (D:x- & G+)[(1-y)ndb]

C: (A- & E+)[nca] or ...
C: (B- & F+)[ncb] or ...

D: (A- & G+)[nda] or (A:C+ & x+)[na]
D: (B- & G+)[ndb] or (B:C+ & x+)[nb]
```

Then, apparently, the counts will be consistent if and only if the same fraction is used when merging connectors.

Connector merging, with counts

To get all of the above correct, there is a series of unit tests. They work well, but one of the more complex ones has become painfully difficult to understand and debug. It is reviewed here. But first, a change of notation to make it more compact:

- The entry `A:(B- & C+)` will be written as `(A, BC)`. Here, the perenthesis denote a pair (a matrix entry). The letter sequence is just the connector sequence with the directional indicators ignored.
- Entries with word-classes, such as `<wclass-AB>:(K- & B+)` will be written as `({AB},KB)`. The word-class is denoted with set-notation curly braces. Similarly, `A:(K- & <wclass-AB>+)` will be written as `(A,K{AB})`.
- Cross-sections, which were written above as `D:(A:C+ & x+)` will be written as `[D, <A, Cv>]`. The angle brackets denote the shape, and the lower-case `v` denotes the location of the variable in the connector sequence. The square brackets just serve to remind that a cross-section is being discussed.
- A property called "detailed balance" is introduced. This is the idea that corresponding sections and cross-sections should have exactly the same observation count on them. Thus for example, given a section `(A, BC)` which was observed N times, one expects that the two cross-sections derived from it, namely `[B, <A, vC>]` and `[C, <A, Bv>]` are also both observed N times each. Prior to any merging, detailed balance holds "automatically" or tautologically, as a trite statement about how counting is done. The goal is that connector merging should preserve detailed balance as a property. It is assumed to be a desirable property, and is enforced in the code and unit tests.
- Counts will be denoted with a lower-case n written in front of the pair. Thus, `n(A, BC)` would be the number of times that `(A, BC)` was observed.

First merge

The troublseome test is '`connector-merge-tricon.scm`'. The relevant portion is as follows. The dictionary is assumed to contain many entries; the troubleseome subset is this:

```
(j, abe)
(f, abe)
```

A decision is made to merge the vectors for `e` and `j`, based on other dictionary entries not shown here. The "projective merge" strategy is used, so that a fraction $0 \leq p \leq 1$ of the count is merged whenever one of the two vectors is missing an entry at a given basis element. In this case, the merge, denoted with an arrow, is

```
none + (j, abe) -> p * ({ej}, abe) + (1-p) * (j, abe)
```

where '`none`' denotes that there is no section `(e, abe)` and so the projective merge was used. That is, the count on `(j, abe)` is reduced to $(1-p)$ of it's earlier value, and the remaining p is transfered over to `({ej}, abe)`. That is, the total counts are preserved. That is,

$$n'(\texttt{\{ej\}, abe}) = pn(\texttt{j, abe})$$

where n' denotes the count after the merge, and the unprimed n is the count before the merge.

From the connector merging discussion above, we conclude that `({ej}, abe)` should be rewritten to `({ej}, ab{ej})`. The count should be as above, that is:

$$n'(\texttt{\{ej\}, ab\{ej\}}) = pn(\texttt{j, abe})$$

The vector on `e` includes the cross-sections

```
[e, <j,abv>]
[e, <f,abv>]
```

These merge in a similar fashion:

```
[e, <j,abv>] + none -> p * [{ej}, <j,abv>] + (1-p) * [e, <j,abv>]
[e, <f,abv>] + none -> p * [{ej}, <f,abv>] + (1-p) * [e, <f,abv>]
```

From detailed balance, we deduce the two new sections

```
(j, ab{ej})
(f, ab{ej})
```

The counts on these are

$$n'(\texttt{j, ab\{ej\}}) = pn(\texttt{j, abe})$$

and

$$n'(\texttt{f, ab\{ej\}}) = pn(\texttt{f, abe})$$

From the connector merging discussion above, we conclude that `(j, ab{ej})` should be rewritten to `({ej}, ab{ej})`. The first identity arrives at the same count as before, so this rewrite appears to be self-consistent. Everything works out.

Second merge

The first merge is more-or-less straightforward. The trouble comes with the second merge. Here, it is decided that the vector for `f` should be merged into `{ej}`. The preservation of detailed balance creates subtleties and ambiguities.

The final count on `({ejf}, ab{ejf})` gets contributions from three sources:

- The starting count on `({ej}, ab{ej})`, which is

  $$n'(\texttt{\{ej\}, ab\{ej\}}) = pn(\texttt{j, abe})$$

  as given above.

- A contribution from `(f, ab{ej})`, which was created in the first merge, via detailed balance from the earlier cross-section `[{ej}, <f, abv>]`. This is merged in it's entirety into the the existing `({ej}, ab{ej})`. The projection merge is

  ```
  ({ej}, ab{ej}) + (f, ab{ej}) -> ({ejf}, ab{ej})
  ```

  This merge absorbs the entire count on `(f, ab{ej})` because `({ej}, ab{ej})` already exists. The contribution is thus $n'(\texttt{f, ab\{ej\}}) = pn(\texttt{f, abe})$. Rewriting then promotes `({ejf}, ab{ej})` to `({ejf}, ab{ejf})`.

- A contribution from `(f, abe)` via the projection merge

  ```
  none + (f, abe) -> q * ({ejf}, abe) + (1-q) * (f, abe)
  ```

  and then the subsequent rewrite of `({ejf}, abe) -> ({ejf}, ab{ejf})`. This contribution is $qn'(\texttt{f, abe}) = q(1-p)n(\texttt{f, abe})$.

The total of these three contributions is then

$$\begin{aligned} n''(\texttt{\{ejf\},ab\{ejf\}}) =& n'(\texttt{\{ej\}, ab\{ej\}}) + n'(\texttt{f, ab\{ej\}}) + qn'(\texttt{f, abe}) \\ =& pn(\texttt{j, abe}) + pn(\texttt{f, abe}) + q(1-p)n(\texttt{f, abe}) \end{aligned}$$

Note that this result is history-dependent: merging j into e first, then f gives a different result than merging f into e, then j (and presumably different than the third possibility, of merging f and j first, and only then adding e).

An open question is whether there is a way of performing the merges that are history-independent, and what would that mean.

Again, additional details are in the test file '`connector-merge-tricon.scm`'.

Connector Merging, Conclusion

After much work: there are ten unit tests, all passing, with the final fix in commit 5e1d7dfb94867f22642d7cdf0621a833bb96092e of 24 May 2021 which fixes a problem not found in the unit tests; it requires a real-world test-case. Need to run `(check-balance LLOBJ)` to evoke it.

expt-19 (May 2021)

Moving on... expt-19 reuses the same corpus as expt-16, and, in order to be comparable to earlier results, reuses the `pair-counts` and the `mpg-parse` disjuncts from expt-16. Here's the dataset statistics, from `print-matrix-summary-report`, without and with shapes:

	cset-only	w/ shapes
rows	12	12
columns	75688	587172
entries	80832	701606
sparsity	3.4901	3.3281
avg. obs./disjunct	12.561	6.4922
entropy	0.7221	5.0270
MMT support	80832	701606
obs. count	3.9e9	15e9

Time to compute the pair-distances was about 0.03 to 1.5 seconds for the cset-only vectors, and 6 to 16 seconds for the shapes. So, at least an order of magnitude slower. That's bad.

The entropy is much much higher, which I interpret as a good thing, indicating that the data is of higher quality.

Using `(gram-classify-greedy-disinfo psa 3.0 4)`, there wern't any merges that got done. That's because similarity never got above 3.0. Here's a table of all positive MI similarities, without shapes, and with. It's sorted on the shape column when the shape column MI is positive, otherwise sorted on the cset-only column.

pair	MI cset-only	MI w/ shapes
!-i	$-\infty$	2.3578
g-i	3.2143	1.8643
a-i	-2.388	1.2513
f-c	1.3032	1.1714
h-c	1.2824	1.0837
f-h	1.2800	1.0757
b-g	0.8982	1.0559
a-WALL	$-\infty$	0.9799
j-b	0.3929	0.5846
j-g	0.3929	0.5265
d-f	0.8141	0.5170
e-b	0.3725	0.5066
d-c	0.7817	0.5026
e-j	0.2484	0.4612
e-g	0.3896	0.4503
d-h	0.8126	0.4446
e-d	-0.320	0.0049
b-i	1.4862	-1.038
a-h	0.6048	-0.079
a-f	0.5596	-0.190
a-c	0.5544	-0.266
d-a	0.5501	-0.257
e-i	0.0819	-1.401
j-i	0.0185	-1.254

In general, any MI of less than 4 is ... pretty distant. Although that statement is true for large-vocabulary systems; here the vocabulary is tiny: 12 words, so I guess an MI greater than 1.0 is actually pretty good.

The distance ‘!-i‘ is alarming. I guess. Seems to indicate that ‘i‘ is usually at the end of sentences. The ‘a-WALL‘ distance suggests that ‘a‘ is frequently at the start of the sentence. If so, this is not obvious from casual inspection of the corpus.

A core problem is that this is a very mixed grammar: lots of ambiguity, lots of word senses, no particular clean factorization. Just looking at it, its quite cloudy as to the actual structure. In particular, although each word belongs to only one word-class, the word-classes have lots of mixed, shared POS entries, and each POS is a seemingly random (duhh) unstructured mess. For example:

- pos-e has single, divalent, trivalent disjuncts on it. Except for a few words, most of English is not like that.
- pos-c has one disjunct. It’s identical to pos-d, which has two disjuncts.
- pos-e has a huge number of disjuncts on it, as do pos-i and pos-j. This seems to allow very grammatically complex sentences to be generated, which “of course” are going to be very hard to decode.

Overall, the grammar appears to be over-complex, and very unlike a natural language grammar. it seems unlikely, just from eyeballing it, that the grammar could be untangled without a very large, exhaustive examination of the corpus. This is a bad experimental base.

Issues:

- Is there any sense in which the word "mixing" is appropriate, in it's technical sense (from ergodic theory?) Can we define mixing from the point of veiw of disjunct ambiguity? Of the indiscernibility of grammars given a corpus?
- Is the automatic grammar generation controlling sufficiently for word-senses? Yes, there's a tunable paramter for that, but some disjuncts accidentally appear in multiple POS, thus making those POS at least partly synonymous.
- Is the current automatic grammar generation API appropriate? It was based on an intuitive sense of factorization, but the randomness seems to eassily generate ambiguous grammars.
- How does one measure the complexity of a grammar?
- Is there some easy way of writing down its factorizability?
- Is there a way of proving that two grammars are equivalent? If two different grammars generate the exact same corpus, then is there some algorithm that can transmute one grammar into the other? How is this found/discovered?
- How can one characterize human natural language grammars? That is, if an artificial grammar is generated, how can we know if it is similar to a natural human language? I don't think the rainbow of human natural languages lines up well (or at all) with the axes of tunable paramters in the grammar generator.

Conclude: the expt-19 grammar is over-complicated, ambiguous, mixed, ugly. We need to restart with a simple grammar.

Also conclude: I do not understand how the ambiguity of grammars works. I do not really understand how factorization works. The artificial grammar generator "works" but I don't understand what it is generating. I don't know how close it is to typical human grammars. There's a bit of a "back to the drawing board" moment here.

expt-20 (May 2021)

Start again, this time with a simple, relatively unambiguous, relatively unmixed grammar. Perhaps even with a tiny artificial subset of English!? Just to make eyeballing easier?

July 2021 - Projective Merge and Entropy Maximization

OK, Above work resulted in a bunch of automation scripts, and a few bug-fixes, a completion of the "shapes" work, but not much in the way of insight. So, restart processing

of English. A rather small corpus reveals a conceptual bug in the projective-merge strategy, when used with mutual information. Projective merge works great (I think .. I guess??) with cosine distance, but not MI. This section describes the problem, and explores solutions.

The small corpus is the "run-2" corpus of the English work, its a truncated copy of "tranche-1" consisting of 3026 articles, 426941 sentences, 8133834 word instances. Word-pair counting went well (except for the handling error that truncated the corpus), as did disjunct formation. The MM^T stats were computed for the disjuncts, including shapes. Everything is fine (nominal) until merge. Projective merge fails dramatically – unrelated words are being merged. Why?

What's the pair-wise MI? The MI here is called the "entropic similarity" in other texts in this directory. It's kind-of the same thing, once one adjusts defintions appropriately. Lets call it MI for short. Lets explore the top seven words that got merged, shown in the table below.

Symmetric-MI							
	of	to	in	he	it	that	said
of	7.460	3.720	4.443	-1.74	-1.44	0.619	-1.43
to		6.848	3.019	0.449	-0.18	0.935	0.070
in			5.699	1.348	0.255	1.517	-0.53
he				3.968	1.854	1.691	1.198
it					2.405	0.819	0.646
that						3.664	0.468
said							3.223

Looks pretty reasonable, right? The self-MI of "it" is shockingly low. Some of the other self-MI's are on the low side, too. But whatever. Small dataset. Asking for an MI > 3.0 for a merge to take place seems like it should provide good results.

The current code base for projective merge recommends a fraction of 0.187 for merging "of" and "to", which seems low, but acceptable. So lets do the merge, by hand (se notes in 'run-2/README' for details. The result is '(WordClassNode "of to")'. What is the MI between this, and the other words? A disaster. See table below.

Symmetric-MI								
	of-to	of	to	in	he	it	that	said
of-to	10.009	8.411	8.216	6.044	1.937	1.629	2.970	1.900
of		11.70	$-\infty$	5.294	-4.44	-3.34	1.012	-3.25
to			11.92	3.558	0.874	-0.25	1.432	-1.97

OK, so what's wrong with that? The naive expectation is that the MI(cluster, "of") and MI(cluster, "to") should be low or negative. It's not - its huge. The naive expectation is that the MI of the cluster to the non-preposition words should remain near zero (small positive or negative) and instead it got higher, not lower! Yow!

Note that the "of" and "to" vectors are the new vectors, with the projected parts removed. Thus, the new MI (of,to)=$-\infty$ is as expected: all overlap is now completely gone. That the MI(of, non-prep) and MI(to, non-prep) has gotten lower is a good sign.

So what's going wrong, here? Well, we failed to define the projective merge in such a way that we minimize and maximize the assorted overlaps. So lets fix that.

Conclusion

The following sections explore the problem above. The answer, in retrospect, appears to be obvious: of the paramterized projective merge, only the overlap merge maximizes entropy. Ading in any fraction of the union merge lowers the total mutual information. It smears out the word-senses, and damages word-sense disambiguation.

If the answer is so obvious, then how did we get into this mess to begin with? Well, it seemed, at the time, that, because the input data is so noisy, that the number of observations is so incomplete, that there might be some advantage for extending the overlap merge with some "small" fraction of the union merge. In retrospect, this appears to have been a failed idea.

The overall concern is still valid: for a small corpus, there is a lot of noise, and perhaps there does need to be some kind of generalization taking place. But perhaps the smearing provided by the union merge is ... not a good idea ... at present. However ...

However, trying to always maximize the entropy runs the risk of stumbling into local maxima, and being trapped by them. This is an old machine-learning problem: how to avoid local maxima. The union merge might eventually be a good way of jumping out of local maxima. However, at this time, I don't understand what's going on clearly enough. It seems to have been a premature feature. Maybe. At any rate, future work should set a baseline of pure overlap merges, zero union merges. If any part of the union is to be mixed in, it should probably be quite small – well less than a percent, rather than the 30% which almost all earlier work made use of.

If the union-fraction is greater than zero, then it should be kept small enough so that the MI between the newly created cluster, and the remainder of the words that were put into the cluster remains less than the cutoff MI that is used to determine if a merge is to be undertaken. Examining the experimental data, below, indicates that this fraction is indeed tiny: its about 0.003 (a third of a percent) for the case that was examined.

Rule of thumb: if there is a merge fraction, then if should be set to

$$f = 1/2^{\max(MI(a,a),MI(b,b))-MI_{cut}}$$

where a,b are the two words to be merged, $MI(a,a)$ is the self-MI and MI_{cut} is the cutoff, below which a merge will not be performed. Any fraction larger than this will results in an MI between the new class, and the remainder of the words a,b that is *larger* than the original $MI(a,b)$ itself ... and that's a disaaster. This is a rule of thumb, because, I guess a more precise value could be given, but getting it requires some more algebra which I'm too lazy to do.

Similarity and Projection

Lets' review the definition for the MI, and the projection.

Entropic Similarity and Cosine Distance

This is a copy of what is in the 'connector-sets-revised.pdf' paper, chapter 6 (pages 39-44). There some additional experimental data in the 'diary-part-one.pdf', pages 102-103.

A word w is associated with a vector $N(w,d_j)$ where the d_j are the disjuncts observed on the word, and N is a count of the number of times that the word-disjunct pair was observed.

For words w,u, define the dot product (inner product) between the words as

$$i(u,w) = \sum_d N(u,d)\,N(w,d)$$

This can be turned into a *bona-fide* joint probability by writing

$$p(u,w) = i(u,w)\,/\,i(*,*)$$

where, as always, * denotes a wild-card – here, a sum over all words:

$$i(*,*) = \sum_{u,w} i(u,w)$$

There is a corresponding marginal probability

$$p(w) = p(w,*) = \frac{i(w,*)}{i(*,*)}$$

The entropic similarity between two words is

$$MI(u,w) = \log_2 \frac{p(u,w)}{p(u)\,p(w)}$$

and written in this form, this is clearly the convetional (fractional) MI between two words.

It's worth comparing this to the cosine distance

$$\theta(u,w) = \arccos \frac{p(u,w)}{\sqrt{p(u,u)\,p(w,w)}}$$

Note that they both start with the dot product in the numerator. The cosine distance is invariant under rotations (orthogonal transformations) in Euclidean space. However, probability space is not Eucliden, so it is a "category error" to work with cosine distance applied to probabilities. By contrast, the MI is "obviously correct" when working with probabilities. In practice, the two are correlated. See page 43 of 'connector-sets-revised.pdf' for experimental data.

The code for computing this stuff can be found in the AtomSpace github repo, https://github.com/opencog/atomspace/blob/master/opencog/matrix/symmetric-mi.scm

To avoid absurdly long compute times, this code uses the MM^T concept (described in connector-sets) to perform and cache partial results that can be quickly combined to obtain the desired MI value. Note, its impossible to pre-compute the MI for any but the smallest vocabularies, as there are just too many words.

Projection Merge

Given two words (word-vectors), the projection merge create three new vectors: a merged vector, and remainders of the two original vectors. This is most easily described in terms of the basis vectors. Define the set of all disjuncts with non-zero counts on word w:

$$D(w) = \{d : N(w,d) \neq 0\}$$

Given a real number $0 \leq f \leq 1$, the merged vector g of the two words w, u has counts

$$N(g,d) = \begin{cases} N(w,d) + N(u,d) & d \in D(w) \cap D(u) \\ fN(w,d) & d \in D(w) \setminus D(u) \\ fN(u,d) & d \in D(u) \setminus D(w) \end{cases}$$

Clearly, the support for g is $D(g) = D(w) \cup D(u)$ whenever $f > 0$ and it is $D(g) = D(w) \cap D(u)$ when $f = 0$. The two new words are u' and w' which are just the old words, with the overlaps removed:

$$N(w',d) = \begin{cases} 0 & d \in D(w) \cap D(u) \\ (1-f)N(w,d) & d \in D(w) \setminus D(u) \end{cases}$$

and likewise for $N(u',d)$.

This projection merge is designed to preserve the total count:

$$N(g,d) + N(w',d) + N(u',d) = N(w,d) + N(u,d) \tag{1}$$

This equation can be called "detailed balance", as it is in thermodynamics.

The merged vector g is meant to corrspond to a "grammatical class"; it is a vector just as any other word-vector, but it is meant to capture the "average" of the two words it is made out of. The role of the fraction f is to handle the situation of the words u and w having multiple word-senses. The idea here is that the set $D(w) \cap D(u)$ captures those disjuncts for which both u and w have teh same word-sense (for example, the parts of u and w that are nouns) whereas $D(w) \setminus D(u)$ and $D(u) \setminus D(w)$ are the disjuncts that belong to other word-senses (*e.g.* the parts of u and w that are verbs). Due to inadequate statistics and systemic noise, a non-zero f might help in smoothing out erroneous assignments of disjuncts to word-senses. In the example above, a value of $f = 0.187$ was used.

The projection described above is the same as that described and implemented in https://github.com/opencog/learn/blob/master/scm/gram-projective.scm (as of this writing).

Maximum Mutual Information

Well, obviously as the example above demonstrates, there's something not quite right with the projection merge, at least, when one works with mutual information.

The projection merge does seem to make sense (mostly) when one thinks of vectors that inhabit Euclidean space: One can easily define the sum of two vectors; and to deal with the fact that the word vectors are themselves sums of multiple senses, one can try to project back out the parts that don't have shared support. The problem with working in Euclidean space is that the counts can go negative: the inner product on Euclidean space really is the cosine (dot) product, and simple linear algebra sense some vector components negative. This is undesirable, as it prevents counts from being interpreted as statistical frequencies.

For probabilities, the correct goal is to maximize the mutual information (maximize the entropy). There seem to be two ways to define the total MI of the system. One is to write

$$MI_{tot} = \sum_{u,w} p(u,w)\, MI(u,w)$$

Note that since both p and MI are symmetric, this has the effect of double-counting the off-diagonal entries. This seems to underweight the diagonal. Another possibility is to double-weight the diagonal:

$$MI_{alt} = \sum_{u \geq w} p(u,w)\, MI(u,w)$$

Not clear, right now, which is better or more correct.

A bit of articulation helps clarify things. From the defintion of MI, one has

$$\begin{aligned}
MI_{tot} =& \sum_{u,w} p(u,w)\, MI(u,w) \\
=& \sum_{u,w} p(u,w) \log_2 \frac{p(u,w)}{p(u)\, p(w)} \\
=& \sum_{u,w} p(u,w) \log_2 p(u,w) - \sum_{u,w} p(u,w) \log_2 p(u) - \sum_{u,w} p(u,w) \log_2 p(w) \\
=& \sum_{u,w} p(u,w) \log_2 p(u,w) - 2 \sum_{w} p(w) \log_2 p(w) \\
=& H_{joint} - 2H_{marg}
\end{aligned}$$

where

$$H_{marg} = \sum_{w} p(w) \log_2 p(w)$$

is the "marginal entropy".

Detailed balance

Detailed balance, eqn 1 means that the merge affects only the rows and columns of the merged words. This is somehow "intuitively obvious", but I'll belabor the topic here,

to make sure we're not making any mistakes. In the following, let a, b be the two words to be merged, creating a category g.

Lemma: For $w \neq a, b$, the row and column sums are unaffected by the merge. That is, one has $i(w,*) = i'(w,*)$.

Proof: Let a, b be the two words to be merged. Then for $w \neq a, b$, one has

$$\begin{aligned} i(w,a) + i(w,b) &= \sum_d N(w,d)\left[N(a,d) + N(b,d)\right] \\ &= \sum_d N(w,d)\left[N(g,d) + N\left(a',d\right) + N\left(b',d\right)\right] \\ &= i(w,g) + i\left(w,a'\right) + i\left(w,b'\right) \end{aligned}$$

No other sums are affected, so that, for $u, w \neq a, b$ one has

$$i(w,u) = i'(w,u)$$

Therefore,

$$\begin{aligned} i(w,*) &= \sum_u i(w,u) \\ &= i(w,a) + i(w,b) + \sum_{u \neq a,b} i(w,u) \\ &= i(w,g) + i\left(w,a'\right) + i\left(w,b'\right) + \sum_{u \neq a,b} i(w,u) \\ &= i'(w,*) \end{aligned}$$

where the i' sum includes (runs over) a', b' and g. □

This enables a key theorem.

Theorem: The total count $i(*,*)$ is unaffected by the merge. That is, $i'(*,*) = i(*,*)$ where the prime denotes the post-merge sum.

Proof: Split out the affected rows and columns. First, the corner case:

$$\begin{aligned} i(a,a) + i(a,b) + i(b,a) + i(b,b) &= \sum_d \left[N(a,d) + N(b,d)\right]\left[N(a,d) + N(b,d)\right] \\ &= \sum_d \left[N(g,d) + N\left(a',d\right) + N\left(b',d\right)\right]\left[N(g,d) + N\left(a',d\right) + N\left(b',d\right)\right] \\ &= i(g,g) + i\left(g,a'\right) + i\left(g,b'\right) + i\left(a',g\right) + i\left(a',a'\right) + i\left(b',g\right) + i\left(b',b'\right) \\ &= i(g,g) + i\left(a',a'\right) + i\left(b',b'\right) + 2i\left(g,a'\right) + 2i\left(g,b'\right) \end{aligned}$$

Two of the terms vanish: $i\left(a',b'\right) = i\left(b',a'\right) = 0$, which is arrived at by noting that the merge has arranged that $N\left(a',d\right) = N\left(b',d\right) = 0$ for $d \in D(a) \cap D(b)$.

Even if these terms did not vanish, one still arrives at

$$i(a,*) + i(b,*) = i(g,*) + i\left(a',*\right) + i\left(b',*\right)$$

Plugging through,

$$\begin{aligned} i(*,*) &= \sum_u i(u,*) \\ &= i(a,*) + i(b,*) + \sum_{u \neq a,b} i(u,*) \\ &= i(g,*) + i(a',*) + i(b',*) + \sum_{u \neq a,b} i(u,*) \\ &= \sum_{u'} i(u',*) \\ &= i'(*,*) \end{aligned}$$

Phew. That was complicated, given that the result seems obvious. □

Corrollary: Likewise, for the probabilities: if a,b are the two words to be merged into g,then $p(a,*) + p(b,*) = p(g,*) + p(a',*) + p(b',*)$.

Proof: Divide by $i(*,*)$. □

Corrollary: The marginal probability of rows/columns not being merged is unaffected by the merge. That is, $p(w) = p'(w)$ for $w \neq a,b$.

Proof:

$$p(w) = p(w,*) = \frac{i(w,*)}{i(*,*)} = \frac{i'(w,*)}{i'(*,*)} = p'(w)$$

Follows from the lemma and the theorem above. □

Corrollary: The mutual information of rows/columns not being merged is unaffected by the merge. That is, $MI(u,w) = MI'(u,w)$ for $u,w \neq a,b$.

Proof:

$$\begin{aligned} MI(u,w) &= \log_2 \frac{p(u,w)}{p(u)\,p(w)} \\ &= \log_2 \frac{p'(u,w)}{p'(u)\,p'(w)} = MI'(u,w) \end{aligned}$$

Follows as before. □

From the above, it is clear that the MI_{tot} splits into an nvariant part, and a part affected by the merge. Write

$$\begin{aligned} MI_{tot} &= \frac{1}{2} \sum_{u,w} p(u,w)\, MI(u,w) \\ &= \frac{1}{2} \sum_{u,w=a,b} p(u,w)\, MI(u,w) + \frac{1}{2} \sum_{u,w \neq a,b} p(u,w)\, MI(u,w) \\ &= MI_{merge} + MI_{invariant} \end{aligned}$$

The focus is then on how MI_{merge} changes. The change in the mutual information due to merging is captured by the difference, defined as

$$S = \sum_{u,w=g,a',b'} p(u,w)\, MI(u,w) - \sum_{u,w=a,b} p(u,w)\, MI(u,w) \tag{2}$$

A suitable name for this would be the "relative entropy", I guess. Seems reasonable.

Experimental Exploration

The goal of later sections will be to find the extrema (maxima, minima) of eqn 2 by algebraic means: that is, to find a value for the parameter f that yeilds the projective merge that maximizes the mutual information. As it turns out, a whole lot of rather tedious algebra is required. Thus, its time for an experimental interlude. Given the dataset above, we can vary the merge parameter by hand, and see what happens.

First, a review os the dataset and the computational techniques. The dataset is 'r2-mpg-trim-40-8-5.rdb'. It contains 6029 words and 98102 disjuncts, and a total of 171922 word-disjunct pairs that were observed a total of 4006152 times. That is, each word/disjunct pair was observed an average of 23.302 times. The dataset is very sparse, with a log2 sparsity of 11.748. This is a trimmed dataset: all words with less than 40 observations were discared; all disjuncts with less than 8 observations were discarded, and all pairs with less than 5 observations were discarded. See the experiment README file for more info on row and column support, average lengths, etc.

Here are the probabilities and MI, before merge:

word-pair	$p(u,w)$	$MI(u,w)$
of,of	1.199e-3	7.460
of,to	7.599e-5	3.720
to,to	5.623e-4	6.848

The total probability is then 1.913e-3 and the weighted MI is 1.336e-2 and the averge MI is then 6.9829. To repeat:

$$\sum_{u,w=a,b} p(u,w) = 1.913 \times 10^{-3}$$

$$\sum_{u,w=a,b} p(u,w)\, MI(u,w) = 1.336 \times 10^{-2}$$

$$\frac{\sum_{u,w=a,b} p(u,w)\, MI(u,w)}{\sum_{u,w=a,b} p(u,w)} = 6.9829$$

for $a,b = of,to$.

Lets repeat this calculation, after merging, with $f = 0.18715$ as before.

pair	$p(u,w)$	$MI(u,w)$
of,of	4.831E-4	11.695
to,to	3.153E-4	11.923
g,g	7.465E-4	10.009
g,of	1.105E-4	8.4107
g,to	7.208E-5	8.2156
total	1.910E-3	10.431

The MI reported in the total column is the paramter-dependent entropy from above:

$$S' = \sum_{u,w=g,a',b'} p(u,w)\, MI(u,w)$$

where a' is what's left of the vector for the word "of", after projection-merging the common part. Likewise, b' is what's left of "to". It is numerically confirmed that a' and b' are orthogonal to each-other: that $p(a',b') = 0$ after the merge. This is as expected — the common components have been merged. That $p(g,a') \neq 0$ is not surprising: an additional portion of a was merged into g, beyond what is strictly possible with the overlap merge. That is, g and a' are intentionally not orthogonal. That's what this is all about – finding out if something broader than a pure overlap merge is somehow advantageous.

Annoyingly, the detailed balance seems to be a bit off — 1.910e-3 *vs.* 1.913e-3 before. That's a fairly hefty amount of rounding error. Surprisingly large... There are 7620 sections and 7620 shapes being merged; calculations should be double-precision, so this rounding error is irritating.

Anyway, that's the baseline. Lets try a range of merge fractions. Again, to reiterate: $f = 0$ corresponds to a pure overlap projection merge: the cluster g consists of only those disjuncts that are shared in common by a and b. Setting $f = 1$ corresponds to the union merge: all disjuncts on a and b are moved into g so that a' and b' are empty after the merge. Other fractions $0 < f < 1$ interpolate linearly between the overlap merge and the union merge. The table below shows what happens.

frac	total MI	MI(g,g)	MI(g,of)	MI(g,to)	p(of,of)	p(g,of)
0.0	11.312	10.499	$-\infty$	$-\infty$	7.312E-4	0
1e-6	11.312	10.499	-8.816	-9.011	7.311E-4	7.27E-10
1e-5	11.311	10.499	-5.494	-5.689	7.311E-4	7.267E-9
1e-4	11.310	10.498	-2.173	-2.368	7.310E-4	7.266E-8
0.001	11.297	10.495	1.1478	0.9527	7.297E-4	7.260E-7
0.002	11.285	10.492	2.1461	1.9510	7.282E-4	1.450E-6
0.005	11.253	10.482	3.4630	3.2679	7.239E-4	3.615E-6
0.01	11.207	10.465	4.4546	4.2595	7.166E-4	7.194E-6
0.02	11.129	10.433	5.4379	5.2428	7.022E-4	1.424E-5
0.05	10.943	10.340	6.7110	6.5159	6.599E-4	3.452E-5
0.1	10.714	10.203	7.6330	7.4379	5.922E-4	6.540E-5
0.187	10.431	10.009	8.4107	8.2156	4.831E-4	1.105E-4
0.3	10.184	9.8344	8.9426	8.7474	3.583E-4	1.526E-4
0.4	10.030	9.7367	9.2374	9.0423	2.632E-4	1.744E-4
0.7	9.7714	9.6468	9.7346	9.5394	6.580E-5	1.526E-4
0.95	9.7000	9.6849	9.9595	9.7644	1.828E-6	3.452E-5
0.99	9.6978	9.6950	9.9873	9.7922	7.312E-8	7.194E-6
1.0	9.6976	9.6976	$-\infty$	$-\infty$	0	0

So, this numeric exploration of the post-merge mutual information reveals ... something that perhaps should have been obvious, in retrospect. Sigh.

In retrospect, it should be clear that the union merge serves only to make word-sense disambiguation cloudier and mushier, by placing unrelated disjuncts into the same class/cluster. This seemed somewhat harmless at some conceptual level, by employing the argumnet that the observed disjuncts are rather noisy, and that perhaps there are commonalities that simply were not observed, due to insufficient sampling. This may still be true, but, viewed from the MI angle, it is clear that anything beyond a pure overlap merge is harmful to the entropy maximization.

Deriving the Entropy Extrema

The goal of this section is to find the extrema of the relative entropy as given in eqn 2. This requires a lot of algebraic calculation. The calculations below were undertaken before the numeric exploration immediately above. The numerics show that this effort was ... pointless. The entropy is maximized by doing the overlap merge, setting $f = 0$; it's it, that's that. The below is simply not necessary. It's kept below for, uhh, posterity. Otherwise, its useless and can be skipped over.

Merge parameterization

It's worth considering the most general merge that maintains detailed balance, and symmetry between the two merged words. This is done by defining a merge vector - a merge parameter for each disjunct, so that one gets a vector $0 \leq f(d) \leq 1$. The merged class is then

$$N(g,d) = \begin{cases} f(d)[N(a,d)+N(b,d)] & d \in D(a) \cap D(b) \\ f(d)N(a,d) & d \in D(a) \setminus D(b) \\ f(d)N(b,d) & d \in D(b) \setminus D(a) \end{cases}$$

The above tries to distinguish the overlapping regions. But it is overly complicted, since on the non-overlapping regions, the counts vanish. That is, $N(b,d) = 0$ when $d \in D(a) \setminus D(b)$. Thus, its enough to write

$$N(g,d) = f(d)[N(a,d)+N(b,d)]$$

To maintain detailed balance as before, one must have

$$N(a',d) = (1-f(d))N(a,d)$$

Note that the number of disjuncts is huge: in practice, a vocabulary of a few thousands words might have hundreds of thousands of disjuncts.[2] The vector $f(d)$ is a large vector.

In what follows, this full vector will be retreated from, going back to the original vision of a projection merge. There are two reasons for this:

[2]The 'run-2' dataset used in the example above had 6029 distinct words and 98102 distinct disjuncts. This was a "trimmed" dataset, 40-8-5: all words with less than 40 observations were discarded. All disjuncts with less than 8 observations were discarded, and all (word,disjunct) pairs with less than 5 observations were discarded. Trimming tends to filter out a lot of the noise in the dataset, without much compromising the data integrity. That is, assuming one can accurately guess where the noise floor is; this remains an open topic.

- The expressions get hopelessly non-linear and intractable,
- It is hard to imagine what sort of additional information might arrive, that would distinguish one disjunct from another. Either a disjunct is associated with a or with b or with both; there really aren't any other possibilities.

However, until that point, a vector of independent f's will be assumed.

Variational problem

A change of notation is in order. Parenthesis have been useful to emphasize what depends on what. Parenthesis are also convenient for plain-ASCII source code documentation. However, for the following, subscript notation will improve readability. Let

$$\begin{aligned} f_d &\equiv f(d) \\ N_{wd} &\equiv N(w,d) \\ M_{uw} &\equiv MI(u,w) \end{aligned}$$

The variational problem is then to obtain

$$\frac{\partial S}{\partial f_d}$$

I see no easy way out, other than to brute-force this. First, note that

$$\begin{aligned} \frac{\partial}{\partial f_d} \sum_{u,w=a,b} p(u,w)\, MI(u,w) &= \frac{\partial}{\partial f_d} \left[p(a,a)\, MI(a,a) + 2p(a,b)\, MI(a,b) + p(b,b)\, MI(b,b) \right] \\ &= 0 \end{aligned}$$

since none of these have $f(d)$ appearning in the expressions. Lets blast away.

$$\frac{\partial S}{\partial f_d} = \sum_{u,w=g,a',b'} M_{uw} \frac{\partial p_{uw}}{\partial f_d} + p_{uw} \frac{\partial M_{uw}}{\partial f_d}$$

So

$$\begin{aligned} \frac{\partial p_{gg}}{\partial f_d} &= \frac{1}{i_{**}} \frac{\partial i_{gg}}{\partial f_d} \\ &= \frac{1}{i_{**}} \frac{\partial}{\partial f_d} \sum_d N_{gd} N_{gd} \\ &= \frac{2}{i_{**}} f_d \left(N_{ad} + N_{bd} \right)^2 \end{aligned}$$

and

$$\begin{aligned}\frac{\partial p_{ga'}}{\partial f_d} &= \frac{1}{i_{**}} \frac{\partial}{\partial f_d} N_{gd} N_{a'd} \\ &= \frac{1}{i_{**}} \left(N_{ad} + N_{bd}\right) N_{ad} \frac{\partial}{\partial f_d} f_d \left(1 - f_d\right) \\ &= \frac{1}{i_{**}} \left(N_{ad} + N_{bd}\right) N_{ad} \left(1 - 2 f_d\right)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial p_{a'a'}}{\partial f_d} &= \frac{1}{i_{**}} \frac{\partial}{\partial f_d} N_{a'd} N_{a'd} \\ &= \frac{N_{ad}^2}{i_{**}} \frac{\partial}{\partial f_d} \left(1 - f_d\right)^2 \\ &= -2 \frac{N_{ad}^2}{i_{**}} \left(1 - f_d\right)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial M_{uw}}{\partial f_d} &= \frac{\partial}{\partial f_d} \log_2 \frac{p_{uw}}{p_u p_w} \\ &= \frac{1}{\log 2} \left[\frac{1}{p_{uw}} \frac{\partial p_{uw}}{\partial f_d} - \frac{1}{p_u} \frac{\partial p_u}{\partial f_d} - \frac{1}{p_w} \frac{\partial p_w}{\partial f_d} \right]\end{aligned}$$

so

$$\begin{aligned}\frac{\partial p_g}{\partial f_d} &= \frac{1}{i_{**}} \frac{\partial i_{g*}}{\partial f_d} \\ &= \frac{1}{i_{**}} \frac{\partial}{\partial f_d} \sum_d N_{gd} N_{*d} \\ &= \frac{1}{i_{**}} \frac{\partial}{\partial f_d} N_{gd} N_{*d} \\ &= \frac{1}{i_{**}} N_{*d} \left[N_{ad} + N_{bd}\right]\end{aligned}$$

and

$$\begin{aligned}\frac{\partial p_{a'}}{\partial f_d} &= \frac{1}{i_{**}} \frac{\partial i_{a'*}}{\partial f_d} \\ &= -\frac{1}{i_{**}} N_{*d} N_{ad}\end{aligned}$$

Putting all this together is painfully tedious. But it must be done. (Is there some

easier way I don't see yet?)

$$\begin{aligned}
\sum_{u,w=g,a',b'} p_{uw}\frac{\partial M_{uw}}{\partial f_d} =& p_{gg}\frac{\partial M_{gg}}{\partial f_d} + p_{a'a'}\frac{\partial M_{a'a'}}{\partial f_d} + p_{b'b'}\frac{\partial M_{b'b'}}{\partial f_d} \\
&+ 2p_{ga'}\frac{\partial M_{ga'}}{\partial f_d} + 2p_{gb'}\frac{\partial M_{gb'}}{\partial f_d} \\
=& \frac{1}{\log 2}\times \\
& \frac{2}{i_{**}} f_d \left(N_{ad}+N_{bd}\right)^2 - 2\frac{p_{gg}}{p_g}\cdot\frac{1}{i_{**}} N_{*d}\left(N_{ad}+N_{bd}\right) \\
& - 2\frac{N_{ad}^2}{i_{**}}\left(1-f_d\right) - 2\frac{p_{a'a'}}{p_{a'}}\cdot\left(-\frac{1}{i_{**}} N_{*d}N_{ad}\right) \\
& - 2\frac{N_{bd}^2}{i_{**}}\left(1-f_d\right) + 2\frac{p_{b'b'}}{p_{b'}}\cdot\left(\frac{1}{i_{**}} N_{*d}N_{bd}\right) \\
& + 2\frac{1}{i_{**}}\left(N_{ad}+N_{bd}\right)N_{ad}\left(1-2f_d\right) + 2\frac{p_{ga'}}{p_{a'}}\cdot\frac{1}{i_{**}} N_{*d}N_{ad} + 2\frac{p_{ga'}}{p_{b'}}\cdot\frac{1}{i_{**}} N_{*d}N_{bd} \\
& + 2\frac{1}{i_{**}}\left(N_{ad}+N_{bd}\right)N_{bd}\left(1-2f_d\right) + 2\frac{p_{gb'}}{p_{a'}}\cdot\frac{1}{i_{**}} N_{*d}N_{ad} + 2\frac{p_{gb'}}{p_{b'}}\cdot\frac{1}{i_{**}} N_{*d}N_{bd}
\end{aligned}$$

Lets hope there are no mistakes. Gathering terms,

$$\begin{aligned}
\frac{i_{**}}{2}\log 2 \sum_{u,w=g,a',b'} p_{uw}\frac{\partial M_{uw}}{\partial f_d} =& f_d\left[\left(N_{ad}+N_{bd}\right)^2 + N_{ad}^2 + N_{bd}^2 - 2\left(N_{ad}+N_{bd}\right)^2\right] \\
& - \frac{p_{gg}}{p_g} N_{*d}\left(N_{ad}+N_{bd}\right) + \frac{p_{a'a'}}{p_{a'}} N_{*d}N_{ad} + \frac{p_{b'b'}}{p_{b'}} N_{*d}N_{bd} \\
& - N_{ad}^2 - N_{bd}^2 \\
& + \left(N_{ad}+N_{bd}\right)^2 \\
& + \frac{p_{ga'}}{p_{a'}} N_{*d}N_{ad} + \frac{p_{ga'}}{p_{b'}} N_{*d}N_{bd} \\
& + \frac{p_{gb'}}{p_{a'}} N_{*d}N_{ad} + \frac{p_{gb'}}{p_{b'}} N_{*d}N_{bd} \\
=& 2\left(1-f_d\right)N_{ad}N_{bd} \\
& + N_{*d}N_{ad}\left[-\frac{p_{gg}}{p_g} + \frac{p_{a'a'}+p_{ga'}+p_{gb'}}{p_{a'}}\right] \\
& + N_{*d}N_{bd}\left[-\frac{p_{gg}}{p_g} + \frac{p_{b'b'}+p_{ga'}+p_{gb'}}{p_{b'}}\right]
\end{aligned}$$

Huf. This appears to be quadratic rational in f because

$$\begin{aligned}
p_{gg} =& \frac{1}{i_{**}}\sum_d N_{gd}N_{gd} \\
=& \frac{1}{i_{**}}\sum_d f_d^2\left(N_{ad}+N_{bd}\right)^2
\end{aligned}$$

while

$$\begin{aligned} p_g &= \frac{1}{i_{**}} \sum_d N_{gd} N_{*d} \\ &= \frac{1}{i_{**}} \sum_d f_d \left(N_{ad} + N_{bd} \right) N_{*d} \end{aligned}$$

So those fractions are nasty. There seem to be two choices in front of us. These are

- Set $f_d = 1$ for $d \in D(a) \cap D(b)$ and $f_d = f$ otherwise. This is the projection merge.
- Set $f_d = f$ for all d. This seems to be pointless; it smears together multiple word senses.

Abandon the second option. To disentangle the location of f in the various quantities, some addtional notation is needed.

The End

This is the end of the diary.