Ambiguity in the UMLS Metathesaurus

2013 Edition

Sonya E. Shooshan, François-Michel Lang and Alan R. Aronson

June 4, 2013

1. Introduction

The UMLS® Metathesaurus® contains a significant amount of ambiguity. For example, the string "Cold" (or "cold" or "COLD") occurs in six distinct concepts with six distinct meanings. The purpose of this report is to examine ambiguity in the 2008AA release of the Metathesaurus in the context of its effect on natural language processing (NLP) applications.

Until the 2004AC release of the UMLS Knowledge Sources, ambiguity was denoted explicitly by appending an ambiguity designator, a number in angle brackets, to the end of an ambiguous string. Thus the ambiguity for "cold" was denoted by 'Cold <1>', 'Cold <2>', 'COLD <3>', etc. Now ambiguity is computed by finding concepts with strings that differ only with respect to case.[1]

Table 1 shows that the degree of Metathesaurus ambiguity has grown over the years and was particularly explosive in 2005, partly due to the direct computation of ambiguity mentioned above.

	2004	**2005**	**2006**	**2007**	**2008**	**2009**
Strings with an ambiguity designator	21,295 (+30%)	N/A	N/A	N/A	N/A	N/A
Concepts with one or more ambiguity	16,775 (+35%)	36,133 (+115%)	44,591 (+23%)	48,820 (+9%)	61,873 (+27%)	71,127 (+15%)
Concepts with one or more non-suppressible ambiguity	12,387 (+19%)	33,513 (+171%)	40,977 (+22%)	43,499 (+6%)	55,168 (+27%)	64,322 (+17%)
Cases of ambiguity	10,018 (+39%)	22,218 (+122%)	27,599 (+24%)	29,415 (+7%)	40,574 (+38%)	45,540 (+12%)
Cases of non-suppressible ambiguity	9,521 (+40%)	20,996 (+121%)	25,290 (+20%)	26,084 (+3%)	36,266 (+39%)	40,937 (+13%)

Table 1. **Measures of ambiguity in the UMLS Metathesaurus**

1. Note that AMBIGSUI.RRF or AMBIG.SUI cannot be used for this purpose because they do not conflate case.

	2010	2011	2012	2013	2014	2015
Concepts with one or more ambiguity	80,999 (+14%)	84,097 (+4%)	87,734 (+4%)	91,629 (+4%)		
Concepts with one or more non-suppressible ambiguity	68,935 (+7%)	75,777 (+10%)	79,357 (+5%)	82,378 (+4%)		
Cases of ambiguity	52,122 (+14%)	57,988 (+11%)	60,764 (+5%)	63,905 (+5%)		
Cases of non-suppressible ambiguity	45,074 (+10%)	51,761 (+15%)	54,353 (+5%)	56,617 (+4%)		

Note that in the table, percentage changes are computed relative to the previous year. More recently, ambiguity grew significantly in 2006 and 2008, less so in 2009 and quite modestly in 2007.

Examining the cases of ambiguity more closely, consider the *degree* of ambiguity, i.e., the number of ways a string is ambiguous or, equivalently, the number of concepts in which it (or one of its case variants) occurs.[1] For example "deprecated ^ wbc-acnc" has degree 124 in 2008 all of which are marked as suppressible; "other" has degree 89 (43 if suppressibles are ignored). Table 2 contains the distribution of ambiguities in the Metathesaurus according to degree. Note that an ambiguity of degree one is not actually an ambiguity. In 2004 and before, for example, 'Abbreviations <1>' is not ambiguous since there were no other 'Abbreviations <n>' strings in the Metathesaurus.

Ignoring suppressible synonyms produces the more realistic distribution shown in Table 4. Most of the ambiguity of higher degree has disappeared, and all of that would disappear if appropriate strings were marked as suppressible. Suppressible synonyms are ignored for the remainder of this report.

Section 2 of this report describes general classes of ambiguity found in the Metathesaurus. Section 3 describes only the most notable cases of ambiguity in the Metathesaurus, i.e., the cases of degree 10 or more. The bulk of the cases are now reported automatically by the Migration Assistant, a tool developed generally for annotating ambiguity and specifically for the purpose of marking appropriate cases as suppressible. Finally, Section 4 is an appendix containing instructions for populating the tables in the report.

1. The computation of the degree of an ambiguity was corrected in 2002. As a result, there are some differences from previous editions of this report in the counts reported in the tables.

Deg of ambig	2006 cases	2007 cases	2008 cases	2009 cases	2010 cases
124		1	1 (0%)		
108					1
93		1			
92	1				
90					1
64					1
54					
89			1	1 (0%)	
39	1	1 (0%)	1 (0%)	1 (0%)	1 (0%)
36	1	1 (0%)	1 (0%)	3 (+200%)	1 (-67%)
25				1	1 (0%)
24	1			3	1 (-67%)
23		1	1 (0%)	3 (+200%)	2 (-33%)
22				1	
21				6	2 (-67%)
20	1		1 (0%)	3 (+200%)	
19	1	1 (0%)		3	1 (-67%)
18	1 (0%)	2 (+100%)	2 (0%)	3 (+50%)	3 (0%)
17			2 (0%)	5 (+150%)	2 (-60%)
16	2 (+100%)	1 (-50%)	1 (0%)	2 (+100%)	3 (+50%)
15	1	3 (+200%)	2 (-33%)	10 (+400%)	4 (-40%)
14	1		3 (+200%)	2 (-33%)	9 (+450%)
13	1	1 (0%)	3 (+200%)	9 (+200%)	7 (-22%)
12	1 (0%)	3 (+200%)	6 (+100%)	12 (+100%)	21 (+75%)
11	3	4 (+33%)	10 (+150%)	13 (+30%)	19 (+53%)
10	4	7 (+75%)	17 (+143%)	18 (+6%)	10 (-44%)
9	13 (+117%)	14 (+8%)	25 (+79%)	40 (+60%)	43 (+8%)
8	23 (+130%)	24 (+4%)	61 (+154%)	70 (+15%)	78 (+11%)
7	28 (+155%)	42 (+50%)	70 (+67%)	118 (+69%)	124 (+5%)
6	66 (+175%)	104 (+58%)	185 (78%)	242 (+31%)	283 (+17%)
5	158 (+193%)	195 (+23%)	404 (+107%)	464 (+15%)	602 (+30%)
4	452 (+117%)	562 (+24%)	996 (77%)	1,231 (24%)	1,360 (+10%)
3	1,868 (+51%)	2,380 (+27%)	4,226 (+78%)	4,873 (+15%)	5,618 (+15%)
2	24,971 (+21%)	26,067 (+4%)	34,555 (+32%)	38,403 (+11%)	43,899 (+14%)
1					
Total	**27,599 (+24%)**	**29,415 (+7%)**	**40,574 (+38%)**	**45,540 (+12%)**	**52,122 (+14%)**

Table 2. **Metathesaurus ambiguity distribution by degree**

Deg of ambig	2011 cases	2012 cases	2013 cases	2014 cases	2015 cases
124					
108					
93					
92					
90	1 (0%)	1 (0%)			
64					
54					
91			1		
89					
39	1 (0%)	1 (0%)	1 (0%)		
38		1	1 (0%)		
36	1 (0%)				
25					
24					
23	2 (0%)	1 (-50%)	1 (0%)		
22					
21					
20			1		
19	1 (0%)	1 (0%)	1 (0%)		
18	2 (-33%)	2 (0%)	1 (-50%)		
17	2 (0%)	1 (-50%)	4 (+300%)		
16	2 (-33%)	6 (+200%)	4 (-33%)		
15	6 (+50%)	3 (-50%)	4 (+66%)		
14	5 (-44%)	4 (-20%)	8 (+100%)		
13	7 (0%)	9 (+28%)	13 (+44%)		
12	11 (-48%)	12 (+9%)	17 (+42%)		
11	13 (-32%)	13 (0%)	18 (+38%)		
10	22 (+120%)	24 (+9%)	27 (+13%)		
9	47 (+9%)	50 (+6%)	65 (+30%)		
8	65 (-19%)	70 (+8%)	93 (+33%)		
7	134 (+8%)	143 (+7%)	159 (+11%)		
6	255 (-10%)	292 (+15%)	321 (+10%)		
5	607 (+1%)	632 (+4%)	662 (+5%)		
4	1,379 (+1%)	1,461 (+18%)	1,583 (+8%)		
3	6,046 (+8%)	6,335 (+4%)	6,745 (+6%)		
2	49,379 (+12%)	51,702 (5%)	54,174 (+5%)		
1					
Total	**57,988 (11%)**	**60,764 (+5%)**	**63,905 (+5%)**		

Table 3. **Metathesaurus ambiguity distribution by degree**

Degree of ambiguity	2006 cases	2007 cases	2008 cases	2009 cases	2010 cases
44					1
43			1	1 (0%)	
41		1			
40	1				
39	1				
36	1	1 (0%)	1 (0%)	3 (+200%)	1 (-67%)
25				1	
24	1			2	
23		1	1 (0%)	4 (+300%)	2 (-50%)
22				1	
21				6	
20	1		1 (0%)	3 (+200%)	
19	1	1 (0%)		3	1 (-67%)
18	1 (0%)	2 (+100%)	2 (0%)	3 (+50%)	1 (-67%)
17				3	
16				1	
15	1	1 (0%)	1 (0%)	9 (+800%)	4 (-55%)
14		1	4 (+300%)	2 (-50%)	5 (+250%)
13	1		1	8 (+700%)	3 (-60%)
12	1 (0%)	3 (+200%)	6 (+100%)	9 (+50%)	12 (+25%)
11	1	2 (+100%)	7 (+250%)	12 (+71%)	10 (-17%)
10	4	6 (+50%)	16 (+167%)	18 (+13%)	20 (+11%)
9	9 (+80%)	12 (+33%)	22 (+83%)	27 (+23%)	28 (+4%)
8	16 (+100%)	19 (+19%)	40 (+110%)	56 (+40%)	51 (-9%)
7	16 (+220%)	25 (+56%)	60 (+140%)	99 (+65%)	87 (-12%)
6	39 (+457%)	87 (+123%)	142 (+63%)	214 (+51%)	202 (-6%)
5	123 (+297%)	160 (+30%)	306 (+91%)	355 (+16%)	424 (+19%)
4	360 (+131%)	481 (+34%)	899 (+87%)	1,133 (+26%)	1,143 (+1%)
3	1,586 (+59%)	2,076 (+31%)	3,857 (+86%)	4,474 (+16%)	4,903 (+10%)
2	23,126 (+17%)	23,205 (+0%)	30,899 (+33%)	34,490 (+12%)	38,175 (+10)
1					
Total	**25,290 (+20%)**	**26,084 (+3%)**	**36,266 (+39%)**	**40,937 (+13%)**	**45,074 (+10%)**

Table 4. **Metathesaurus ambiguity distribution after removing suppressibles**

2. Classes of Metathesaurus Ambiguity

Some concepts contain strings which should be marked as suppressible. Many of these strings are already marked suppressible for a given UMLS release; this report recommends further cases

Degree of ambiguity	2011 cases	2012 cases	2013 cases	2009 cases	2015 cases
45			1		
44	1 (0%)	1 (0%)			
43					
41					
40					
39					
38		1	1 (0%)		
36	1 (0%)				
25					
24					
23	2 (0%)	1 (-50%)	1 (0%)		
22					
21					
20					
19	1 (0%)	1 (0%)	1 (0%)		
18	2 (+100%)	2 (0%)	1 (0%)		
17			2		
16	1	3 (+200%)	3 (0%)		
15	15 (0%)	5 (-66%)	4 (-20%)		
14	7 (+40%)	4 (-42%)	3 (+33%)		
13	13 (0%)	4 (-69%)	4 (0%)		
12	7 (-42%)	8 (+14%)	10 (+25%)		
11	16 (+60%)	15 (-6%)	18 (+20%)		
10	15 (-20%)	16 (+7%)	15 (-6%)		
9	33 (+18%)	38 (+15%)	33 (-13%)		
8	54 (+6%)	56 (+4%)	70 (+25%)		
7	106 (+22%)	120 (+13%)	129 (+8%)		
6	221 (+9%)	258 (+17%)	278 (+8%)		
5	505 (+19%)	508 (0%)	528 (+4%)		
4	1,265 (+11%)	1,345 (+6%)	1,420 (+6%)		
3	5,572 (+15%)	5,852 (+5%)	6,205 (6%)		
2	43,949 (+15%)	46,115 (+5%)	47,877 (+4)		
Total	**51,761 (+15%)**	**54,535 (+5%)**	**56617 (+4%)**		

Table 5. **Metathesaurus ambiguity distribution after removing suppressibles**

some of which are universally applicable and some of which are appropriate in more limited environments such as the natural language processing done by MetaMap.

The analysis in this and previous editions of this report reveals some classes of ambiguity commonly occurring in the Metathesaurus:

- **Contextual (or hierarchical) ambiguity.** This class of false ambiguity is exemplified by the string 'prostate' for 'Prostatic Diseases'. (Many of these problems have been fixed by suppressing the misleading string for the concept; but the problems continue to reappear as the Metathesaurus grows.) It normally arises from terms which require context within their vocabulary (in this case, a disease hierarchy) in order to be properly understood. Contextual ambiguities can be classified according to their participants:
  - **Body part/disease ambiguity** exemplified by 'Prostate' and 'Prostatic Diseases'
  - **Body part/procedure ambiguity** exemplified by 'Stomach' and 'Procedures on the stomach'
  - **Pathology/procedure ambiguity** exemplified by 'Pathology' and 'Pathology procedure'
  - **Medical device/procedure ambiguity** exemplified by 'Prosthesis' and 'Prosthesis Implantation'
  - **Substance/therapy ambiguity** exemplified by 'Anthracyclines' and 'prior anthracycline therapy'
  - **Substance/measurement ambiguity** exemplified by 'Thyroid stimulating immunoglobulins (TSI)' and 'Thyroid stimulating immunoglobulins assay'
- **Generalization ambiguity.** This is also false ambiguity caused by grouping several concepts together using a more general term. For example, 23 concepts including 'Protocols: Activities' and 'Protocols: Pre- or Intra- or Post-Procedure**'** are generalized to 'Protocols' which does seem to be a legitimate synonym of the concept 'Protocols documentation'.
- **Meta ambiguity.** This new class of ambiguity, represented by strings such as 'Stress fracture, NEC in ICD10_1998', contain meta information. In this case it is the name of the vocabulary, ICD10_1998 in the example. As opposed to the first class of ambiguity above in which strings such as 'Prostate' meaning 'Prostatic Diseases' do not say enough about themselves, these strings say too much. It is true that the meaning of a string containing 'NEC', 'not elsewhere classified' or like phrase, depends upon its vocabulary, but such information is already available in the MSRO file (where it belongs). It is also true that such strings have different meanings and strictly speaking should be different concepts. But the practical result of such a representational scheme is to introduce an ambiguity that most users do not want or need to resolve. (It is not even clear that those who might want to resolve the ambiguity can do so with the information available in the Metathesaurus.)
- **Abbreviation ambiguity.** This is another, large class of ambiguity caused by distinct concepts having the same acronyms (or abbreviations). An example from above is that 'Mitral Valve Stenosis', 'Multiple Sclerosis', 'Morphine Sulfate' and 'millisecond' all have abbreviation 'MS' or 'ms'. Although this class represents true ambiguity in a strict sense, it is better to disallow it in many text processing situations, especially those in which authors define the abbreviations they use. Unlike the other classes of ambiguity defined above, we do not recommend that this case be reflected in changes to the Metathesaurus. This kind of ambiguity will be suppressed for MetaMap processing only.

3. Higher Degree Metathesaurus Ambiguity

Ambiguous English Metathesaurus strings are described in this section in decreasing order of degree of ambiguity. Only those cases of degree 10 or more are covered. See Migration Assistant reports for cases of ambiguity of lesser degree.

In all cases, suppressible synonyms are ignored as is done in Table 4. Ambiguous forms for concepts shown in bold should be marked as suppressible. Recommendations for cases which are not clear are introduced with the word *consider*. Ambiguous forms for concepts shown in italics should be marked as suppressible in MetaMap only.

These data are no longer being collected and updated. The most recent version of the report containing these data is 2011.

4. Appendix

Data contained in all tables in this report are obtained from the current year's ambiguity study directory, `$NLS/specialist/module/metawordindex/data.*/01Ambiguity`.

4.1 Populating Table 1

The following data are available in the `0log3.*` file:

1. For concepts with one or more ambiguity:
   `wc -l ambiguity_cases.cuis`
2. For concepts with one or more non-suppressible ambiguity:
   `wc -l supp.ambiguity_cases.cuis`
3. For cases of ambiguity:
   `wc -l ambiguity_cases.unique`
4. For cases of non-suppressible ambiguity:
   `wc -l supp.ambiguity_cases.unique`

4.2 Populating Tables 2 and 3

To populate Table 2 simply fill in the values, adding new rows as necessary, from the file `ambiguity_cases.counts` in the ambiguity study directory; to populate Table 4 use the file supp.`ambiguity_cases.counts` instead; both these tables are captured in the `0log3.*` file