Ambiguity in the UMLS Metathesaurus

2010 Edition

**Francois-Michel Lang, Sonya E. Shooshan,
James G. Mork and Alan R. Aronson**

October1, 2009

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 each 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%)					
Concepts with one or more non-suppressible ambiguity	68,935 (+7%)					
Cases of ambiguity	52,122 (+14%)					
Cases of non-suppressible ambiguity	45,074 (+10%)					

The numbers in the chart in this report are obtained from the 0log3 logfile in the directory

/nfsvol/nls/specialist/module/metawordindex/data.NN/01Ambiguity

where NN represents current (two-digit) year.

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 2010, 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 3. 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.

Degree of ambiguity	2007 cases	2008 cases	2009 cases	2010 cases
124	1	1 (0%)		
108				1
93	1			
92				
90				1
64				1
54				
89		1	1 (0%)	
39	1 (0%)	1 (0%)	1 (0%)	1 (0%)
36	1 (0%)	1 (0%)	3 (+200%)	1 (-67%)
25			1	1 (0%)
24			3	1 (-67%)
23	1	1 (0%)	3 (+200%)	2 (-33%)
22			1	
21			6	2 (-67%)
20		1 (0%)	3 (+200%)	
19	1 (0%)		3	1 (-67%)
18	2 (+100%)	2 (0%)	3 (+50%)	3 (0%)
17		2 (0%)	5 (+150%)	2 (-60%)
16	1 (-50%)	1 (0%)	2 (+100%)	3 (+50%)
15	3 (+200%)	2 (-33%)	10 (+400%)	6 (-40%)
14		3 (+200%)	2 (-33%)	9 (+350%)
13	1 (0%)	3 (+200%)	9 (+200%)	7 (-22%)
12	3 (+200%)	6 (+100%)	12 (+100%)	21 (+75%)
11	4 (+33%)	10 (+150%)	13 (+30%)	19 (+46%)
10	7 (+75%)	17 (+143%)	18 (+6%)	33 (+83%)
9	14 (+8%)	25 (+79%)	40 (+60%)	43 (+8%)
8	24 (+4%)	61 (+154%)	70 (+15%)	78 (+11%)
7	42 (+50%)	70 (+67%)	118 (+69%)	124 (+5%)
6	104 (+58%)	185 (78%)	242 (+31%)	283 (+17%)
5	195 (+23%)	404 (+107%)	464 (+15%)	602 (+30%)
4	562 (+24%)	996 (77%)	1,231 (24%)	1,360 (+10%)
3	2,380 (+27%)	4,226 (+78%)	4,873 (+15%)	5,618 (+15%)
2	26,067 (+4%)	34,555 (+32%)	38,403 (+11%)	43,899 (+14%)
1				
Total	**29,415 (+7%)**	**40,574 (+38%)**	**45,540 (+12%)**	**52,122 (+14%)**

Degree of ambiguity	2007 cases	2008 cases	2009 cases	2010 cases
44				1
43		1	1 (0%)	
41	1			
40				
39				
36	1 (0%)	1 (0%)	3 (+200%)	1 (-67%)
25			1	
24			2	
23	1	1 (0%)	4 (+300%)	2 (-50%)
22			1	
21			6	
20		1 (0%)	3 (+200%)	
19	1 (0%)		3	1 (-67%)
18	2 (+100%)	2 (0%)	3 (+50%)	2 (-33%)
17			3	
16			1	
15	1 (0%)	1 (0%)	9 (+800%)	4 (-55%)
14	1	4 (+300%)	2 (-50%)	5 (+150%)
13		1	8 (+700%)	3 (-63%)
12	3 (+200%)	6 (+100%)	9 (+50%)	12 (+25%)
11	2 (+100%)	7 (+250%)	12 (+71%)	10 (-17%)
10	6 (+50%)	16 (+167%)	18 (+13%)	20 (+11%)
9	12 (+33%)	22 (+83%)	27 (+23%)	28 (+4%)
8	19 (+19%)	40 (+110%)	56 (+40%)	51 (-9%)
7	25 (+56%)	60 (+140%)	99 (+65%)	87 (-12%)
6	87 (+123%)	142 (+63%)	214 (+51%)	206 (-4%)
5	160 (+30%)	306 (+91%)	355 (+16%)	424 (+19%)
4	481 (+34%)	899 (+87%)	1,133 (+26%)	1,143 (+1%)
3	2,076 (+31%)	3,857 (+86%)	4,474 (+16%)	4,903 (+10%)
2	23,205 (+0%)	30,899 (+33%)	34,490 (+12%)	38,175 (|11%)
1				
Total	**26,084 (+3%)**	**36,266 (+39%)**	**40,937 (+13%)**	**45,074 (+10%)**

Table 3. **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 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 3. 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.

The following table identifies all ambiguous English Metathesaurus strings whose degree of ambiguity is at least 10.

Degree	String
44	other
36	unknown
23	leber optic atrophy
23	protocols
19	assessment
18	ec 2.7.1.112
18	patient education plans
15	ar
15	ec 2.1.1.43
15	emergency
15	pap
14	cap
14	cd
14	ms
14	none
14	t
13	active
13	alp
13	cat
12	a
12	as
12	c
12	ec 1.14.14.1
12	ec 2.7.1.-
12	genotropin kit
12	m
12	normal
12	not applicable
12	p
12	p40
12	ptc
11	ad
11	asp
11	cam
11	car
11	gas
11	p14
11	patient
11	tr

Degree	String
11	tw
11	yes
10	adh
10	at3
10	bar
10	cad
10	ec 3.4.22.-
10	f
10	ice
10	k
10	kit
10	mac
10	mycorrhizal samples
10	no
10	psa
10	psp
10	pt
10	radiology
10	sports medicine
10	t1
10	u
10	y