ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/329173608

Computer-assisted checking of conceptual relationships in a large thesaurus: Proceedings of the Fifteenth International ISKO Conference 9-11 July 2018 Porto, Portugal

Chapter · January 2018

DOI: 10.5771/9783956504211-128

CITATIONS
0

READS
245

4 authors:

Decio Wey Berti Junior
Federal University of Minas Gerais
5 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Gercina Angela de Lima
Federal University of Minas Gerais
142 PUBLICATIONS **558** CITATIONS
SEE PROFILE

Benildes Coura Moreira dos Santos Maculan
Federal University of Minas Gerais
132 PUBLICATIONS **205** CITATIONS
SEE PROFILE

Dagobert Soergel
University of Maryland, College Park
259 PUBLICATIONS **3,679** CITATIONS
SEE PROFILE

All content following this page was uploaded by Benildes Coura Moreira dos Santos Maculan on 31 January 2019.

The user has requested enhancement of the downloaded file.

Decio Wey Berti, Jr., Gercina Lima, Benildes Maculan and Dagobert Soergel

Computer-assisted checking of conceptual relationships in a large thesaurus

Abstract

We describe a method to support quality control of relationship instances in a large thesaurus or other KOS, using the example of AGROVOC (~33K concepts and ~97K conceptual relationship instances), where manually checking each relationship instance is not feasible. Our method identifies relationship instances that should be checked manually; it can also shed light on problems with the definition of relationship types. We apply a simplified version of the linguistic concept of verb valency to the analysis of conceptual relationships, treating relationship types as verbs. We map each of the two concepts in a relationship instance to an entity type; the resulting entity type pair is a valency pattern, as in the following example:

Flavivirus < *causes*> yellow fever ➡ Valency pattern [microorganism, diseaseOrDisorder

A relationship instance that use a valency pattern that is rare for the relationship type might be erroneous and should be checked by an editor. We describe our method in detail, how we associated concepts with the appropriate entity type (this information is not available for AGROVOC) and how we organized the data for analysis. Then we present some illustrative results.

1. Introduction. The problem

Relationships between concepts form the skeleton of thesauri and other Knowledge Organization Systems (KOS). They are of enormous importance for the use of KOS to support retrieval and as knowledge bases for artificial intelligence applications. Many thesauri use only very broad conceptual relationships, hierarchical (BT/NT) and associative (RT). There has long been a call for refined relationship types (Schmitz-Esser, 1999; Soergel *et al.* 2004). Large thesauri have tens of thousands of relationship instances, thus introducing refined relationships requires huge effort. Relationships are useful only if they are of high quality. This paper addresses the issue of quality control in the establishment of relationships between concepts given that in most cases checking all relationship instances manually would be prohibitively expensive.

We use AGROVOC as our test environment because it is a large thesaurus and uses refined relationships, the focus of this paper. AGROVOC has ~33K concepts, and ~97K conceptual relationship instances; of these, ~35K use BT; ~35K use NT; ~4K RT, and~ 23K use refined relationship types listed in the *AGROVOC Ontology*.

2. Literature review and conceptual background

2.1. Literature review

This paper is in the general area of finding errors in large KOS automatically. We saw three approaches.

Approach 1 uses purely formal checks (unprintable characters) to more content-

related checks (duplicate preferred labels, missing scope notes, or issues in the relational structure). The qSKOS program by Mader 2017 is a good example.

Approach 2 uses use statistical analysis of text *corpora* to find, for example, instances of problematic equivalence between a term E in English and a term P in Portuguese. Using an English *corpus* and a Portuguese *corpus* on the same topic, one can check the occurrence patterns of E and P in their respective *corpus*; if E and P mean the same, then the occurrence patterns should be similar (Nohama *et al.* 2012)

Approach 3 analyzes relationship instances using the entity types (semantic types) of the concepts that are connected. Mougin and Bodenreiter 2008 analyze the consistency of relationships in the NCI Thesaurus with relationships in the UMLS Semantic Network, but they apply this method only to derive a global measure of consistency, not to find individual errors in relationship instances. Jiang, Solbrig, and Chute 2012 also use UMLS semantic types to find errors in a specific type of KOS, a list of common data elements (CDE), such as *Dosage Unit of Measure Code*, in medical records with their associated permissible values. All permissible values must belong to the same semantic type, otherwise there is an error. In the example, the permissible value *capsule* is an error; *capsule* is not a dosage unit. The first study uses a top-down approach, starting with the UMLS Semantic Network. The second study's method is similar to ours but in a very specific and simple context. We use a bottom-up approach that starts with analyzing relationship instances in AGROVOC to find atypical relationship instances that should be checked by an editor; we could not find prior work that uses this method.

2.2. Conceptual background

We apply a simplified version of the linguistic concept of verb valency (Perini 2015) to the analysis of conceptual relationships. Summarizing from Perini: A verb may occur in one or more grammatical constructions or *syntactic-semantic-schemata*. Such a schema, also called a *valency pattern*, specifies the syntactic and semantic (or thematic) roles of a verb's complements and possibly the types of concepts that can fill these semantic roles. For example, consider the construction or valency pattern

[6]	VSubj > *Agent*	V	NP > *Patient*	with example
[7]	The cook	melted	the cheese	
[ ]	VSubj > *Agent*	V	NP > *Recipient*	NP > *Theme*
[ ]	Jim gave	his girlfriend	a cake	

Linguists study the valency patterns associated with a verb in a text *corpus*. The set of all valency patterns associated with a verb is the verb's valency.

In our approach to the analysis of relationships in a thesaurus or other KOS, we treat the relationship types as verbs. Most KOS are restricted to binary relationships (often represented as RDF triples); so the constructions are Concept1 V Concept2. Each concept has a semantic role and a concept type or *entity type*. We use a much simplified

form of valency patterns, a pair of entity types, [entityType1, entityType2].

3. Methods

3.1. Methods, general principle

Our method is based on checking for each relationship instance the entity types of the concepts connected by the relationship. In the simplified world of binary relationships, such a pair of entity types constitutes a *valency pattern*, see Table 1.

Table 1: Valency patterns

Valency pattern	Relationship instance example
[namedPlaceOrLocation, namedPlaceOrLocation]	Canada <*spatiallyIncludes*> Fraser River
[microorganism, diseaseOrDisorder	Flavivirus < *causes*> yellow fever

For each relationship instance we derive a valency pattern by mapping each of the two concepts to an entity type. This enables two types of analysis:

1. For a relationship type, list the associated valency patterns by frequency. The occurrence of several frequent valency patterns suggests that the meaning of the relationship should be examined. Infrequent valency patterns suggest that the corresponding relationship instances should be examined; the relationship may be used incorrectly. This is the aspect we are focusing on in this paper (Table 2).
2. For a valency pattern, list all associated relationships by frequency. This will shed light on the interpretation of the relationship types and may suggest some realignment of the definition and use of relationship types (Table 3).

Table 2: Analysis of a relationship type by associated valency patterns

	Relationship type: <*spatiallyIncludes*>		
	Valency pattern	**Relationship instance example**	**Freq**
1 good	[namedPlaceOrLoc, namedPlaceOrLoc]	Canada **<*spatiallyIncl*>** Fraser River	537
2 bad	[physiographic Feature, physiogrFeature]	lowland **<*spatiallyIncl*>** valleys	3
3 bad	[physiogrFeature, namedPlaceOrLoc]	Boreal forests **<*spatiallyIncl*>** Arctic tundra	2
4 bad	[namedPlaceOrLoc, organization]	Canada **<*spatiallyIncl*>** IDRC (International Development Research Centre)	1

The example in row 1 clearly makes sense; the relationship <*spatiallyIncludes*> can exist only between individual physical objects; AGROVOC further restricts the use of <*spatiallyIncludes*> to entities of type namedPlaceOrLocation. We can now turn our attention to the relationship instances that use low-frequency valency patterns. To make

sense of row 2, the relationship could be re-interpreted to apply to universals as well, in the sense that each namedPlaceOrLocation that is of type lowland <*spatiallyIncludes*> a namedPlaceOrLocation of type valley, which is clearly not the case. Row 2 would be a good relationship instance only under the interpretation that a namedPlaceOrLocation that is of type lowland often or sometimes <*spatiallyIncludes*> a namedPlaceOrLocation of type valley. Similar considerations show that row 3 is not a good relationship instance; what the editor wanted to express is that the Arctic tundra is a Boreal forest. The relationship in row 4 does not work at all since IDRC is not a namedPlaceOrLocation; the IDRC building is, so Canada <*spatiallyIncludes*> IDRC building would be ok.

Table 3: Analysis of relationship types associated with a given valency pattern

	Valency pattern [namedPlaceOrLocation, namedPlaceOrLocation]		
	Relationship type	**Relationship instance example**	**Freq**
1a 1b 1c	<*spatiallyIncludes*>	Canada <*spatiallyIncludes*> Fraser River Argentina < *spatiallyIncludes*> Falkland Islands tropical America < *spatiallyIncludes*> Brazil	537
2a bad 2b 2c 2d bad	<*includes*>	Madagascar <*includes*> Mangoky River United Kingdom <*includes*> Falkland Islands USA <*includes*> Guam Latin America <*includes*> Central America	46
3a* 3b bad 3c* 3d*	<*hasMember*>	Francophone Africa <*hasMember*> Mauritania Latin America <*hasMember*> Brazil OECD countries <*hasMember*> United Kingdom Small Island Developing States <*hasMember*> Belize	212
4 bad	<*hasPart*>	USA <*hasPart*> Puerto Rico	3

From this table one can make many interesting observations and identify problems; we list here just a few. Rows 2a and 2d should be <*spatiallyIncludes*>. It appears that one meaning of <*includes*> in this context refers to political inclusion. 3b should be <*spatiallyIncludes.*> In 3a, c, and d we have on the left side groups (or sets) of named places; there should be a separate entity type for these (perhaps an issue with our approximate method of entity types, see Section 3.2). <*hasMember*> then makes sense if one uses a very loose definition. In AGROVOC, <*hasMember*> is also used for membership in organizations, such as ASEAN, which is a much more formal relationship. As seen from the prevailing examples, row 4 should be <*includes*>.

3.2. Methods, implementation

We needed to assign entity types to AGROVOC concepts so we could determine the valency pattern used in a relationship instance, accomplished through Steps 1 – 3.

1. A computer program constructed a tree-structure hierarchy of AGROVOC

concepts starting from *AGROVOC Top Concepts* following NT relationships down (Fig. 1).

2. Starting from the *Basic Formal Ontology (BFO)* class hierarchy, we developed a hierarchy of entity types specifically for AGROVOC by examining the AGROVOC hierarchy. (Fig. 2).
3. We then manually assigned entity types to concepts by taking advantage of the hierarchy: We identified segments of the hierarchy (some large, some small) so that all or most concepts in the segment belonged to the same entity types, as can be seen from Fig. 1. This assignment is approximate and contains some errors.
4. We also arranged the relationship types used by AGROVOC into our own hierarchy to support analysis (Figure 3).
5. We mapped concepts to their entity types and created a massive table of relationship instances to facilitate analysis (Fig. 4).

Figure 1: AGROVOC hierarchy pieces**.**

entities [AGROVOC sense]
. world [for us: namedPlaceOrLocation]
. . continents
. . . Americas
. . . . North America
. . . . . Canada.

organisms
. microorganisms
. . viruses
. . . Flaviviridae
. . . . Flavivirus

features
. physiographic features
. . land cover
. . . vegetation
. . . . forests
. . . . . forest types (by species)
. . . . . . coniferous forests
. . . . . . Boreal forests

phenomena
. biological phenomena
. . disorders
. . diseases
. . physiological functions
. . . respiration

Figure 2: Entity type hierarchy pieces

E_1 BFO:continuant
E_1.1 BFO:independentContinuant
E_1.1.1 BFO:materialEntity
E_1.1.1.1 BFO:object
E_1.1.1.1.1 inanimateObject
E_1.1.1.1.2 animateObjectOrganism
E_1.1.1.1.3 bodyPart
E_1.3 BFO:specificallyDependentContinuant
E_1.3.1 BFO:quality == property
E_1.3.2 BFO:realizableEntity
E_1.3.2.0 stateCondition
E_1.3.2.3 BFO:disposition
E_1.3.2.3.1 diseaseOrDisorder
E_1.3.2.3.2 BFO:function

E_2 BFO:occurrent
E_2.1 BFO:processBroad
E_2.1.1 process
E_2.1.1.1 processHappening
E_2.1.1.2 OBI:plannedProcessOrActivity
E_2.1.1.2.1 activity
E_2.1.1.2.2 methodTechnique

E_3 nonBFOEntities
E_3.1 namedPlaceOrLocation
E_3.3 scientificScholarlyArea
E_3.4 workersProfessions

OBI = ***Ontology of Biological Investigations***

Figure 3: Relationship type hierarchy

r_04	**spatialRelations**
r_04.01	**.** surrounds
r_04.02	. spatiallyIncludes
r_05	**$temporal relations**
r_05.01	**.** precedes
r_06	**quantitativeRelationship**
r_06.01	**.** greaterThan
r_06.02	**.** measuredBy
r_06.03	**.** usingValue

r_07	**includes**
r_07.02	**.** hasMember
r_07.03	**.** includesSubprocess
r_08	**hasPart**
r_08.01	**.** hasComponent
r_08.02	**.** isComposedOf
r_08.03	. hasComposition
r_08.04	**.** hasPortion

Figure 4: Data table for analysis, simplified

Relationship type	Entity type 1	Entity value 1	Entity type 2	Entity value 2
spatiallyIncludes	namedPlaceOrLoc.	Canada	namedPlaceOrLoc.	Fraser River
causes	microorganism	Flavivirus	diseaseOrDisorder	yellow fever

The actual table also includes the hierarchically structured notations for the entity types and relationship types, making it easy to "aggregate up" in the analysis.

4. Illustrative results

This section continues Section 3.2 and further illustrates our method at work examining the relationship types <*includes*> and <*surrounds*> with emphasis on shedding light on the definition and general usage of these relationship types.

Table4 shows the top valency patterns for the relationship type<*includes*>.

Table 4: Some top valency patterns for the relationship type <*includes*>

	Valency pattern	Example	Freq
1	[chemSubstance, chemlSubst.]	heavy metals <*includes*> mercury	219
2	[macroorganism, macroorganism]	Decapoda <*includes*> crabs	193
3a 3b 3c	[activity, activity]	sectoral planning <*includes*> agricult. planning home economics <*includes*> cooking risk management <*includes*> risk assessment	80
4	[otherMaterial, otherMaterial.]	soil parent materials <*includes*> rock	65
5	[taxonProperty, macroorganism]	spring crops <*includes*> Triticum	63
6	[object, object]	farm equipment <*includes*> harvesters	56
7	[namedPlaceOrLoc., namedPl.OrLoc.]	USA <*includes*> Guam	46
8	[economicSector , economicSector]	agroindustry <*includes*> fertilizer industry	30
9	[bodyPart , bodyPart]	olfactory organs <*includes*>nose	12

The relationship type <*includes*> has 1,759 relationships instances; it uses 30 valency patterns that occur 10 or more or times, accounting for 1,192 relationship instances, 20 that occur 5-9 times, accounting for 134 relationship instances and 300 that occur 1-4 times, accounting for 433 relationship instances. Some of this variety may be due to errors in our approximate entity type assignments. Here we examine the top 7 valency patterns and two more selected for illustration. In most of the rows in Table 4 the valency patterns the two entity types in the pair are the same. Most of the relationship instances use <*includes*> in the meaning of set inclusion or hierarchy, one of the meanings of "includes" in natural language. This interpretation applies also to row 5; taxonProperty refers to a set of Taxa that have the property. However, natural language "includes" has several meanings (name a member of a group or class; cover; encompass, includes as constituent or part), and this ambiguity carries into the use of <*includes*> in AGROVOC. Row 7 expresses that USA as a political entity has a part Guam. In row 3a agricultural planning is a type of sectoral planning, but in row 3b cooking is a constituent activity of home economics, not a type of home economics, and the same analysis applies to rows 3c and row 8. So perhaps <*includes*> should be split into two relationships, or the hierarchical <*includes*> should be replaces with <*hasMember*> and the constitutive <*includes*> with <*hasPart*>.

Table 5 shows valency patterns for the relationship type <*surrounds*>. This relationship is related to <*spatiallyIncludes*> (mostly used for namedPlaceOrLocation) and <*hasPart*> (used for bodyPart, among others), with an important distinction: The heart is not a part of the pericardium, but it is surrounded by or enclosed in the pericardium. Similarly, the Gulf of Thailand is not part of Thailand but (partially) surround by Thailand. For bodies of water that are completely included in a country, the relationship <s*patiallyIncludes*> is used.

Table 5: Valency patterns for the relationship type <*surrounds*>

	Valency pattern	**Example**	
1	[bodyPart, bodyPart]	pericardium <*surrounds*> heart,	15
2	[namedPlaceOrLocation, namedPlaceOrLocation]	Thailand <*surrounds*> Gulf of Thailand	4
3	[otherMaterial, otherMaterial]	coffee pulp <*surrounds*> coffee beans,	1
4	[otherMaterial ,developmentalStageAgeGroup]	cocoon <*surrounds*> pupae,	1
5	[physiographicFeature, namedPlaceOrLocation,]	Boreal forests<*surrounds*> Arctic region,	1

This table give rise to some general observations to clarify definitions of entity types and relationship types. As with <*spatiallyIncludes*>, the concepts connected with <*surrounds*> should be individuals, as in row 2. To make this work for row 1, we need

to interpret this relationship as: In any individual organism, the pericardium in this organism surrounds the heart in this organism.

Row 3 shows a problem in our entity assignments; coffee pulp and coffee bean should be bodyPart. Similar considerations apply to row 4, but the entity type of pupa and similar concepts that refer to the whole body or a body part at a given stage of development need further thought.

Row 5 is an error.

looking at out large data table, we could make many other observations. To facilitate analysis, we prepared two major arrangements:

1 sorted by relationship type, then valency pattern and

2 sorted by valency pattern and then by relationship type

One observation relevant here is this: <*spatiallyIncludes*> is used for namedPlaceOrLocation, <*hasPart*> is used for BodyPart (among others), <*surrounds*> is used for both. This does not appear entirely consistent.

5. Conclusion

Large thesauri and other KOS that are well-structured using fine-grained relationship types are useful for information retrieval and reasoning in knowledge-based systems. But such KOS are hard to check and clean to assure that relationship instances are correct and then to maintain that high quality. We set out to develop a method that makes quality control feasible by identifying relationship instances that should be checked by an editor. We have demonstrated through numerous examples that our method of applying a simplified version of the linguistic concept of verb valency shows great promise for realistic quality control for large thesauri and other KOS.

Note

To obtain the full data illustrated in Figures 1-4, including the data table with the AGROVOC relationship instances, contact ds@dsoergel.com.

References

AGROVOC. http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus Triples from http://202.45.139.84:10035/catalogs/fao/repositories/agrovoc on 2016-07-15.

Ontology http://aims.fao.org/sites/default/files/uploads/file/aos/agrontology/index.htm *Top Concepts*. http://aims.fao.org/standards/agrovoc/linked-data (scroll down).

BFO. Basic Formal Ontology. http://ifomis.uni-saarland.de/bfo/.

Jiang, G., Solbrig, H. R., & Chute, C. G. (2012). Quality evaluation of value sets from cancer study common data elements using the UMLS semantic groups. *Journal of the American Medical Informatics Association*, *19*(e1), e129-e136. https://doi.org/10.1136/amiajnl-2011-000739.

Mader, C. (2017). *qSKOS: Vocabulary quality assessment tools*. Java. Retrieved from https://github.com/cmader/qSKOS (original work published 2011).

Mougin, F., & Bodenreider, O. (2008). Auditing the NCI thesaurus with semantic web technologies. *AMIA Annual Symposium Proceedings* 2008, p. 500-504. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655981/.

Nohama, P., Pacheco, E. J., Andrade, R. L., Bitencourt, J. L., Markó, K., & Schulz, S. (2012). Quality issues in Thesaurus building. A case study from the medical domain. *Revista Brasileira de Engenharia Biomédica*, *28*(1), 11–22. https://doi.org/10.4322/rbeb.2012.002.

OBI. The Ontology for Biomedical Investigations, PLoS One. 2016 Apr 29;11(4). Also http://obi-ontology.org/.

Perini, M. A. (2015). Describing Verb Valency: Practical and Theoretical Issues. Springer.

Schmitz-Esser, W. (1999) "Thesaurus and beyond: An Advanced formula for linguistic engineering and information retrieval". Knowledge Organization, Vol. 26, No. 1, 10-22.

Soergel, D.; Lauser, B; Liang, A.; Fisseha, F.; Keizer, J.; Katz, S. (2004). Reengineering thesauri for new applications. The AGROVOC example *Journal of Digital Information*, Volume 4 Issue 4, 2004 March, Article No. 257, http://journals.tdl.org/jodi/article/view/112/111.

View publication stats