Polysemy and Synonymy Detection in Ontology Engineering

ARTEMIS CHALEPLIOGLOU[1,2], SOZON PAPAVLASOPOULOS[2] & MARIOS POULOS[2]
[1]Department of Archival, Library and Information Studies,
University of West Attica,
12243 Athens
GREECE
&
[2]Department of Archives, Library Science and Museology,
Faculty of Information Science & Informatics,
Ionian University,
491 00 Corfu,
GREECE

Abstract: - Polysemy, when a single term has multiple meanings, and synonymy, when multiple terms have the same meaning, are common phenomena in linguistics as well as in scientific knowledge. In ontology engineering, it is vital to detect the synonyms annotations and the multiple inheritances because of polysemy. The persistence of these issues in the semantic description of a knowledge domain causes problematic interoperability and data processing. The disambiguation of the entities, properties and relationships sense in a semantic web ontology significantly improves linked data generation and information retrieval. We explore the synonymy and polysemy in the setting of a cardiology terminology generated from textbooks on the basis of field coverage, professionals' associations' recommendations and bibliometrics, for the building of a cardiologic ontology. From 56,134 terms collected we found that 67.7% were unique. The indexed terms included single words, compound words and multi-word expressions. The frequency of their appearances in the combined master index was calculated and used as a marker of their significance. To cope with the linguistic polysemy and synonymy of terms, we examined them in WordNet, MeSH and BioPortal, as well as by latent semantic analysis (LSA) through singular value decomposition (SVD). Through these approaches we managed to identify and decipher semantic associations and relationships between the terms. We proposed a roadmap for ontology building from scratch by utilizing intrinsic and extrinsic knowledge resources and reuse of metadata. We anticipate that this approach is applicable in ontology engineering of different knowledge domains for relationships setting and linked data contextualization.

Key-Words: - Cardiology, Index analysis, Latent Semantic Analysis, Singular Value Decomposition

Received: January 23, 2020. Revised: May 12, 2020. Accepted: July 15, 2020. Published: July 31, 2020.

1 Introduction

To make the most of the web documents, semantic representation is necessary. Semantic web is based upon ontologies and plenty of them have been developed in the past 20 years [1]. Capture source web metadata, the success of ontology building, is strongly linked to the understanding of the knowledge domain described. The definition of entities, the classes, the properties, the function terms and the individuals, as well as the syntax of expression of restrictions, and the axioms logical rules is a complex, multi staging, repetitive and continuously evolving progress. Therefore, the building of new ontologies upon a clean corpus of terms is important for both research and practical reasons.

We recently described the process of selection and formation of cardiological terms to develop an ontology [2]. Cardiology was selected because cardiovascular diseases represent the top non-communicable disease epidemic worldwide, with increasing numbers in morbidity and mortality despite the progress of modern medicine and pharmacology [3]. Previously several similar efforts have been described such as the CardioVascular Research Grid (CVRG) [3, 4], the representation of heart development in the gene ontology [5], the circulatory system ontology based on ICD-11 and

SNOMED CT [6], and the implantable electronic devices recordings ontology [7]. The predominant challenge for the cardiology field remains the accurate representation of the multiplex interplay between clinical, physiological, pathological, pharmaceutical and biological entities. To accurately represent this knowledge domain by collecting all the necessary entities for this task we applied bibliographic reasoning to extract them from cardiology related textbooks according to their frequency of appearances [2]. The meaning of terms should be accurately defined to support relationships and reasoning.

In this work, we propose a roadmap for the building of the basal terminology and relationships scheme to build a novel ontology, describing a scientific knowledge domain. We anticipate that the disambiguation of terms will significantly facilitate computer reasoning.

2 Problem Formulation

In biomedicine multi-words expressions (MWEs) are used to describe exact anatomic locations, physiological conditions, biological molecular entities and mechanisms with such accuracy to allow one-to-one correlations or hierarchical relationships. For instance, “heart failure” a well described clinical condition with unique MeSH ID changes meanings with the introduction of a single word in the multi-word expression of the term (Fig.1). The term becomes more specific into expressing clinical conditions, speed of progression, aetiology of disease or a specific anatomical entity. While the “heart failure” paradigm could be straightforwardly represented and resolved with “is a kind of” relationships, linguistic hypernnyms, hyponyms and co-hyponyms, other cases are far more complicated. The “broken heart” term represents such a complex case with different meanings in different settings. The “broken heart” is commonly defined as devastating sorrow and despair, a feeling rather than a pathological cardiology condition. But in the setting of cardiology the “broken heart” stands for Takotsubo cardiomyopathy, a syndrome of transient left ventricular apical dysfunction with a high risk of arrhythmia, associated with high levels of catecholamines because of extreme stress (MeSH unique ID: D054549), whilst it could be used as a synonym of “heart failure” (MeSH unique ID: D006333) in a broader sense. In addition, in the Semantic web a word or a MWE may be associated with a distributed semantic representation [8]. Hence, the linguistic determinants of terms should be considered in parallel with the scientific ones and should be recorded, evaluated and incorporated in this ontology to achieve the desirable level of interoperabiblity with other ontologies in the Linked Data ecosystem.

To resolve this, we utilized bibliographic and linguistic methods to decipher semantic relationships in cardiology to represent: (a) clinical (anatomical, physiological and pathological), (b) biological (genes and proteins), and (c) therapeutic (interventions, drugs and devices) modalities. The criterion of the selection of the terms included in the analysis was the frequency of their appearances in cardiology textbooks and their kinship relationships from linguistic and biomedical perspectives. The kinship relationships of terms were explored in three settings: (a) when a term has multiple meanings (polysemy), (b) when multiple terms have the same meaning (synonymy), and (c) when different terms expressed in multi-word expressions (MWEs), analyzed into words, share one or more words in common.

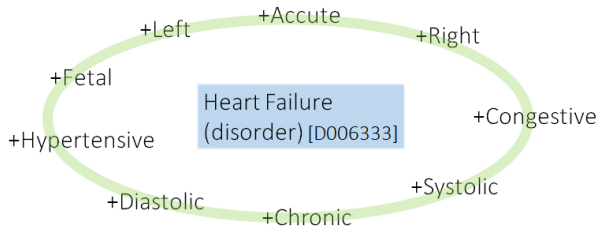

Fig.1 Heart failure as a paradigm of specialization of meaning by introducing a word in multi-word expression terms.

3 Problem Solution

The cardiology knowledge domain was explored in the setting of the terms indexed in a well-defined collection of scientific textbooks describing this field. Briefly, twenty-five textbooks were selected according to the degree of field coverage, the recommendations and guidelines of cardiology professionals’ societies, and their popularity [2]. The textbooks were thematically categorized into general cardiology, pathology physiology and molecular biology ones. The broad thematic specialization of textbooks was necessary to ensure the widest coverage of cardiology field aspects.

3.1 Selection of terms

The index of each textbook was recovered as a set o terms Ai = (a1, a2, a3, …, an), whereas i is the serial number of textbook, between 1 and 25, and a are the terms. Single words and MWEs were included in the indexed terms. The terms were extracted from each book in text format (TXT) and transformed into comma-separated values (CSV) files. The terms

were alphabetically sorted in the ascending order in a single column as a list. These lists were subsequently merged into a new file in a single column without removing duplications. This new list (master index) was sorted alphabetically in the ascending order. This set A = (at1, at2, at3, …, at56134), whereas t is a term, was analyzed for the frequency of appearances of each term according to the logical function counter, if ati = ati-1 then add 1 to the term frequency counter unless ati ≠ ati-1, a condition that terminates the counter and restarts the operation. A new list of terms was generated after the elimination of duplicates as B = (t1, t2, t3, …, t18128) accompanied by their reciprocal counters list in a different column C = (t_c1, t_c2, t_c3, …, t_c18128), whereas t_ci represent the counter of appearances of the term ti in the textbook indices, in a two column formatted table (Fig.2, top left).

The items were arranged in descending order from highest to lowest arithmetic value of the counter of term appearances. We found that the top ten referenced terms in the cardiological textbooks were the anatomical adjective term "ventricular" with 481 appearances, the anatomical term "pulmonary venous" with 338, the treatment noun "management" with 287, the noun and adjective clinical condition "hypertensive" with 234, the diagnostic noun term "electrocardiography" with 229, the general medical noun term "pathophysiology" with 228, the diagnostic noun term "echocardiography" with 223, the medical adjective term "diagnostic" with 215, the medical "treatment with" with 212, noun "cardiomyoplasty" with 204 appearances.

A total of 11786 out of the 18128 unique terms, 65% of the unique terms, appeared only once in the cardiological textbook indices tested, 2590 terms, 14.3%, appeared twice, 1130 terms, 6.2%, appeared thrice, 609 terms, 3.4%, four times, 378 terms, 2.1%, five times, 266 terms, 1.4%, six times, 190 terms, 1.0%, seven times, 159 terms, 0.9%, eight times, 129 terms, 0.7%, nine times, and 98 terms, 0.5%, ten times.

The number of terms versus the number of their appearances in indices exhibits best fit with log-log regression model following the equation:

$$y = a\,x^{\beta} \tag{1}$$

with an R^2=0.998 (α = 18764 and β = 1.318).

By applied the Hirsch index function:

$$h - index = \max_{i} \min(f(i), i) \tag{2}$$

we define that at least 68 out of the 18128 unique cardiological terms, 0.4% of the total, appear 68 or more times in the master index.

Integration of the power function of terms vs their appearances (1) results in:

$$\int \alpha\, x^{\beta}\, dx = \frac{\alpha x^{\beta+1}}{\beta+1} + C \tag{3}$$

By calculating the definite integral of the function (3) for the h-index we found that it represents 70% of the total area under the curve of terms vs their appearances. Thus, it could be postulated that this set of unique terms may sufficiently describe this knowledge field.

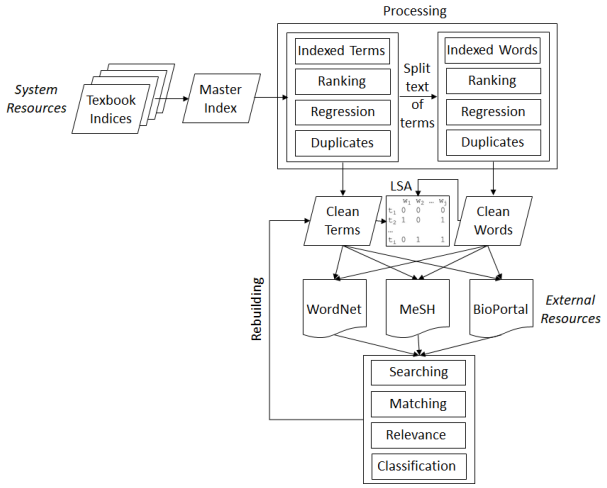

Fig.2 The ontology building architecture applied.

3.2 Analysis of indexed terms in words

The analysis performed allowed the formation of a clean list of terms that can be used for the development of a cardiological ontology. To cope with the linguistic and semantic phenomena of term definition, polysemy and synonymy, we explored the terms into three external resources WordNet, MeSH and BioPortal (Fig.2, middle).

WordNet, a Princeton University initiative, was selected as a comprehensive English linguistic tool, with a lexical database including nouns, verbs, adjectives and adverbs organized into sets of cognitive synonyms (synsets). WordNet interlinks both word forms as well as specific senses of words by identify the meaning of the queries [9]. Looking for the top 500 terms identified in the master index results into a multicolumn table with the information on the type of word, the number of its senses, and its kinship relationships, in specific coordinate terms, hypernyms, hyponyms, meronyms, and antonyms. MIT Media Lab ConceptNet could be considered as an alternative in capturing senses of concepts and relationships, however it is not a linguistic tool and it lacks of specific scientific terms. ConceptNet may be used for common terms but such terms are out of the scope of this analysis.

MeSH, the National Library of Medicine Medical Subject Headings thesaurus, browser was

used the standard biomedical terminology. Browsing for the top terms in appearances identified led to the formation of a separate table with unique identifiers, including RDF unique identifier for ontology building, synonyms, tree structures, treetops and preferred concept name [10]. To the best to our knowledge it is the most comprehensive controlled medical vocabulary in English language.

BioPortal, the predominant library of biomedical ontologies, annotator was used to retrieve annotations of each top identified term in biomedical ontologies and collect classes, ontologies, context and matches results [11]. Currently there is no alternative collection of biomedical datasets other than BioPortal.

The collected evidence from the analyses of top terms in external resources was used in matching, relevance checking and classification of them and ultimately into the rebuilding of the terms list by using preferred terms instead of synonyms (Fig.2, bottom and middle).

To cope with MWEs we added a second level in the analysis of the indexed terms (Fig.2, top right) as follows. The indexed terms were split into words with MS Excel split text to columns data tool to the depth of a maximum of ten words. The resulted columns were collected, copied and pasted, and merged into a single column resulting by eliminating the blanks in a list of 144815 single words with repeats. This table E = (aw1, aw2, …, aw144815), whereas aw stands for the words was further analyzed as afore mentioned for the terms of the master index with repeats. Briefly, The words were alphabetically sorted in the ascending order in a single column as a list. The frequency of appearances of each word was calculated utilizing the logical function counter, if awi = awi-1 then add 1 to the word frequency counter unless awi ≠ awi-1, a condition that terminates the counter and restarts the operation. A new list of words was generated after the elimination of duplicates as G = (w1, w2, w3, …, w16516) accompanied by their reciprocal counters list in a different column K = (w_k1, w_k2, w_k3, …, w_k16516), whereas w_ki represents the counter of appearances of the word wi in the master word index list.

The items were arranged in descending order from highest to lowest arithmetic value of the counter of word appearances. We found that 8269 out of the 16416 single words, 50.1% of the total, were mentioned only once in the index, while 2663 words, 16.1%, twice, 1277 words, 7.7%, thrice. The top ten referenced words in the index were “heart” with 1412 appearances, “ventricular” with 1311, “cardiac” with 1237, “disease” with 787, “pulmonary” with 746, “atrial” with 714, “aortic” with 704, “coronary” with 667, “myocardial” with 648 and “risk” with 552 appearances. By ranking, regression and duplicate word elimination a clean words index list was generated. textbook indices, in a two column formatted table (Fig.2, top left).

The unique single words of this table were further explored in the external resources tested, WordNet, MeSH and BioPortal as performed for the terms. The collected information from the analyses of top single words from the external resources was used in matching, relevance checking and classification of them and ultimately into the rebuilding of the terms through the evaluation of terms relationships after crosscheck unique single words with unique MWEs by latent semantic analysis (LSA) examination.

For the LSA analysis we utilized the singular-value decomposition technique [12] and the code for a matrix (word x context) analysis previously described [13] (Fig.2, middle centre). As we see the terms were split into words, in other terms decomposed into a set of 10 vectors. In our approach the MWEs decomposed terms were used instead of documents to identify relevance or exact match with the single words. In specific the top 500 words of the G = (w1, w2, w3, …, w16516) list was cross examined with the top 500 terms from the B = (t1, t2, t3, …, t18128) list in a wi x tj (word x context) fashion. The frequency, fij, was calculated in matrix according to the presence or not of wi word in the tj term. These frequencies were transformed to the first order association of a word and term:

$$\frac{\log(f_{i,j}+1)}{-\sum_{1-j}\left(\left(\frac{f_{i,j}}{\sum_{1-j} f_{i,j}}\right)*\log\left(\frac{f_{i,j}}{\sum_{1-j} f_{i,j}}\right)\right)} \quad (4)$$

We found that 38,005 terms mentioned only once in the master index, 11,786 terms twice, 2,590 trice, whilst five terms appeared more than 350 times.

The matrix was then analyzed by singular value decomposition [ij] = [ik] [kk] [jk]’ where [ij] the occurrence matrix, [iki] and [jk] orthonormal columns, [kk] the diagonal matrix of singular value where k≤max(i,j). Only the largest singular values dimensions are retained. Each word is represented as a vector of length dimensions. Herein, because the terms could be decomposed into a limited number of dimensions depth the analysis presents limitations. However, it was suggested as the closer applicable solution to examine MWEs similarity of meaning in our dataset in detail by overcoming the two major issues of polysemy and synonymy for information retrieval [13]. The length of vectors was

used as a measure of these similarities between different MWE terms. By compare them against single words we identified relationships between terms automatically.

4 Conclusion

The accurate representation of a scientific knowledge domain by ontology depends on the inclusion of the necessary and sufficient terms describing it, as well as by defining their meaning and relationships. No matter how many terms have been included in such a highly specialized ontology the problem of the exact definition of them is critical for the publisher to set relationships, restrictions and rules for reasoning. However, this work is challenging because of linguistic or scientific polysemy and synonymy.

We proposed a five step approach to build a vocabulary and relationship structure for an ontology (Fig.2). Firstly, the publisher should define a system resources dataset, in this case a textbook collection, that sufficiently describe the field. This material was used to extract the keyword terms, including relationship terms, in a master index, whereas each term could be represented multiple times. The number of appearances of each term in the textbooks could be used as a marker of each critical importance in the building of an ontology describing this field.

Therefore, in the second step, first level of analysis, the terms were ranked according to their appearances, followed by regression and elimination of duplicates. Since, MWEs are in the list of terms, it is necessary to add a second level of analysis within the second step. This level is based in term decomposition by split MWEs in single words. This analysis adds several dimensions for each term and generates a new index of singular words with repeats. Then again the number of appearances of a single word could be used as a marker of its significance in the description of the field. As we can see this analysis produces a different ranking of single words-terms when compared to MWE-terms generated ranking. Ultimately, the second step of analysis produces two lists, one of clean terms and one of clean words.

The third step of the analysis is the question of terms and words meaning explored in three levels of information retrieval external resources, covering the linguistic expressions, the knowledge filed thesaurus scientific expressions, and the already published linked data expressions. Through searching, matching, relevance and classification, the third step provides to the publisher the information of meaning and relationships of MWEs and single words. This step also provides evidence on polysemy and synonymy through the comparison, matching and relevance but it depends on author manual reasoning.

To cope with polysemy and synonymy automatically we proposed a forth step, the LSA analysis where single words are explored against MWEs contexts to identify relationships. This approach has the limitation of the limited depth of MWEs decomposition in multiple dimensions because of their relative short number of words. However, although not tested in this setting because of limitations in computing power, LSA could be suggested for the analysis of words against the definitions extracted from the third step of the analysis which allows significantly higher degree of dimensions.

Finally, The outcomes of the third and forth step provide the material to the ontology publisher to reevaluate and rebuild the ontology by keeping, revisiting or introduce terms, relationships, restrictions and rules. The combination of two resources of information for the publisher, an intrinsic, selected, designed and build by him according to his view and understanding of the scientific field, together with independent by him extrinsic resources, covering different aspects of information, general linguistic, specialized scientific and other ontologies satisfies unbiased analysis of the domain. The experience of the filed professionals is also incorporating through this approach. The general linguistic searching and matching provides important information on parent and child terms relationships, while the grammatical form type of words provide the means for natural language processing. A majority of adjectives as single words was observed in our dataset which are commonly used to specify the meaning of a noun term, clinical, biological or pharmacological, in a MWE, by an anatomical location, developmental stage, molecular pathway or chemical modification manner. Future studies will explore the network relationships between MWEs and single words as critical parts of the ontology building components.

The mathematical description of terms as a function of their repeats and of words as a function of their appearances as well as dimensions in LSA of MWEs allow the determination of the best-fitted selection of terms and relationships in an ontology. The proposed roadmap for the development of a novel ontology from scratch on the basis of intrinsic resources, such as textbooks, and extrinsic resources, both linguistic and scientific, for the formation of a terms list, descriptive items and

relationships, which could be updated and reevaluated continuously, is applicable in different information settings and knowledge domains and therefore is of additive value for ontology building, linked data contextualization and information science.

References:

[1] A. Chaleplioglou, S. Papavlasopoulos, M. Poulos, BioPortal Ontologies Integration with SNOMED CT, RxNORM & GO Datasets, *2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO)*, 2019, pp. 170-175.

[2] A. . Chaleplioglou, S. Papavlasopoulos, M. Poulos, Minimization of Terms to Describe a Knowledge Domain for Ontology Engineering and Linked Data Generation, *WSEAS Transactions on Information Science and Applications*, Vol.16, 2019, pp. 64-68.

[3] R. .L. Winslow, J. Saltz, I. Foster, J.J. Carr, Y. Ge, M.I. Miller, L. Younes, D. Geman, S. Graniote, T. Kurc, R. Madduri, T. Ratnanather, J. Larkin, S. Ardekani, T. Brown, A. Klasny, K. Reynolds, M. Shipway, M. Toerper, The CardioVascular Research Grid (CVRG) Project, in: *Proceedings of the AMIA Summit on Translational Bioinformatics 2011*, 2011, pp. 77-81.

[4] S. Steinert-Threlkeld, S. Ardekani, J.L. Mejino, L.T. Detwiler, J.F. Brinkley, M. Halle, R. Kikinis, R.L. Winslow, M.I. Miller, J.T. Ratnanather, Ontological labels for automated location of anatomical shape differences, *Journal of biomedical informatics*, Vol.45, 2012, pp. 522-527.

[5] V.K. Khodiyar, D.P. Hill, D. Howe, T.Z. Berardini, S. Tweedie, P.J. Talmud, R. Breckenridge, S. Bhattarcharya, P. Riley, P. Scambler, R.C. Lovering, The representation of heart development in the gene ontology, *Developmental biology*, Vol.354, 2011, pp. 9-17.

[6] J.M. Rodrigues, S. Schulz, A. Rector, K. Spackman, J. Millar, J. Campbell, B. Ustun, C.G. Chute, H. Solbrig, V. Della Mea, K.B. Persson, ICD-11 and SNOMED CT Common Ontology: circulatory system, *Studies in health technology and informatics*, Vol.205, 2014, pp. 1043-1047.

[7] A. Rosier, P. Mabo, M. Chauvin, A. Burgun, An ontology-based annotation of cardiac implantable electronic devices to detect therapy changes in a national registry, *IEEE journal of biomedical and health informatics*, Vol.19, 2015, pp. 971-978.

[8] F.A. Nielsen, L.K. Hansen, Creating semantic representations. In: S. Sikström, D. Garcia (eds.), *Statistical Semantics*, Springer, 2020.

[9] C. Fellbaum, K. Brown, WordNet and wordnets. In: Brown, Keith et al. (eds.), *Encyclopedia of Language and Linguistics, Second Edition*, Oxford: Elsevier, 2005.

[10] J. Willis, Searching MeSH Treetops, *NLM Technical Bulletin*, Vol.343, 2005, pp. e2.

[11] P L. Whetzel, N.F. Noy, N.H. Shah, P.R. Alexander, C. Nyulas, T. Tudorache, M.A. Musen, BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications, *Nucleic Acids Research*, Vol.39, 2011, pp. W541-W545.

[12] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, Vol.41, 1990, pp. 391-407.

[13] J.I. Maletic, A. Marcus, Using latent semantic analysis to identify similarities in source code to support program understanding, Proceedings *12th IEEE Internationals Conference on Tools with Artificial Intelligence, ICTAI 2000, IEEE*, 2000, pp. 46-53.

[14] Author, Title of the Paper, *International Journal of Science and Technology*, Vol.X, No.X, 200X, pp. XXX-XXX.

[15] Author, *Title of the Book*, Publishing House, 200X.

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Author Contributions: Please, indicate the role and the contribution of each author:

Example

Artemis Chaleplioglou carried out the indexes collection and the analysis of terms. She designed the ontology building architecture and she performed the indexes analysis over WordNet, MeSH, BioPortal and LSA. She was responsible for the Statistics and Mathematics applied and she prepared the current manuscript.

Sozon Papavlasopoulos supervised the project and reviewed the current manuscript.

Marios Poulos supervised the project and reviewed the current manuscript.

.

Follow: www.wseas.org/multimedia/contributor-role-instruction.pdf