Acquisition and Enrichment of a Subject Domain Thesaurus from a Sample of Scientific Abstracts

Alexander Voinov [1]

[1] *Saint-Petersburg Polytechnical University, Polytechnicheskaya St 29, Saint-Petersburg, 195251, Russia*

Abstract

Two approaches to acquire a hierarchical thesaurus of a subject domain, defined by abstracts, resulting from a query to Pubmed are proposed. One extends the MeSH thesaurus with the use of an NLP technique. Same technique, together with methods of qualitative data analysis, is a key to establish an ad hoc semantic hierarchy of terms from the scratch. The proposed method allows to build a balanced tree of clusters, which may serve an initial approximation to an ontology of the subject domain. No apriori ontology or a thesaurus of a wider subject domain is needed. Semantic associations, which are implicit to the corpus of analyzed texts, are elicited via custom methods of multidimensional scaling and cluster analysis.

Keywords

ontology, thesaurus, multidimensional scaling, cluster analysis, deep natural language parsing

1. Introduction

Ontology is a conceptualization tool, which adds layers of abstraction and generalization over sources of knowledge such as natural language texts. Ontology consists of concepts with transitive or non-transitive relations defined over them (has-part, is-a, functions-as etc), together with more complex logical rules and invariants defined over concepts. Internal properties of a concept are usually described as lists of attribute-value pairs. Concepts may have instances, usually specific to a certain domain of application.

A vocabulary of terms of a subject domain is often used as an explicit or implicit foundation for a conceptual structure of that domain. That vocabulary usually is given a hierarchy, which reflects semantic or linguistic links between terms. In contrast to taxonomies of full-fledged ontologies, multiple overlapping hierarchy “trees” can be defined over vocabulary terms, as done, e.g., in MeSH [1].

Many publicly available ontologies of specific subject domains have been built to date (see, e.g., [2] for an overview). Therefore, in practical applications, one could start with one of those ontologies and extend (or enrich) it with concepts, specific to the application purpose ([2]).

Voinov [3] proposes an approach to ontology integration, at which logical and semantic conflicts between sources of knowledge, reflecting different aspects of the same subject domain, are not eliminated, but rather moved to “local” ontologies, which concretize and develop a more general one. That way one gets a set of ontologies as integral objects, interrelated according to a more or less general scope of knowledge which they describe.

To further develop that approach and to test it on practical applications, one has to have access to a large number of related, but not subsuming each other vocabularies or thesauri of subject domains. An approach, described in the current paper, aims at developing a method of generation of a hierarchical thesaurus by means of a deep NLP of textual sources of knowledge, e.g., scientific abstracts. That method should only use publicly available texts and be moderate in terms of consumption of computational and human resources.

Russian Conference on Artificial Intelligence (RCAI-2021), October 11–16, 2021, Taganrog, Russia
EMAIL: avoinov@gmail.com (A. 1)
ORCID:

2. Materials and Methods

Two publicly available textual databases were used at different stages of the project, Pubmed and USA Today. Pubmed, a repository of papers on medicine and life sciences, has been finally given a priority, because a scientific abstract, by its purpose, is much more condensed and focused, in terms of content conveyed, that a newspaper article. In a sense, an abstract can be regarded as a list of keywords, which is given a literature form to facilitate reading. Moreover, almost all Pubmed articles are annotated by MeSH terms. Pros and cons of using MeSH thesaurus in statistical literature mining are given in [4].

An NLP anaylysis is done with the use of publicly available tools and libraries: WordNet; Python packages, such as NLTK and BeatifulSoup; a Brown Corpus [5].

WordNet is a natural choice for a source of enrichment of an ontology with terms, mentioned in texts (both scientific and not). Approaches, based on WordNet, are proposed in [6,7,8]. Use of noun phrases (NPs), extracted from texts, to enrich ontologies is described in [9,10,11].

A way to enrich MeSH thesaurus with terms, extracted from texts of Pubmed abstracts, is described in one of the sections below. However, the main focus of the paper is brought to an original method of building a taxonomy of terms from the scratch, using custom methods of multidimensional scaling and cluster analysis [4,12,13] on top of similarity measures between terms.

3. Parsing Pubmed Abstracts to extract Noun Phrases

A set of Pubmed abstracts is most easily obtained with the use of its query interface. E.g. *'social network [tw] OR social media [tw]'* ([tw] means: use only text words in the text, ignore all other data and metadata). That query, at the moment of writing this text, results in about 29000 articles.

Each of the resulting abstracts (to meet the goals of the study, we skipped those which were less than 140 characters long) is then a subject of NLP parsing with the use of NLTK functions. All Noun Phrases (NPs), which conform to a grammar expression of ***NP: {<JJ>*<NN.*>+}***, are extracted. That is, zero or more adjectives followed by one or more nouns. That way we identify all left definitions of the English languag, which are found in the text, e.g. “efficient real time database design principles” or the like. For other languages the grammar expression may look different but would center around the main word of a substantive noun expression.

For our example query we get the following list of most frequent NPs, which satisfy the grammar expression above:

Table 1.
Stats on NPs in a query result

Noun Phrase	Number of occurrences
social medium	1258
Study	1167
covid-19	776
People	501
….	…
Group	169
Spread	167
social media platform	165

Considering that example, as well as subsequent ones, it should be noted, that Pubmed is focused on medicine and related domains of knowledge. As a result, relatively neutral topics, such as “social media”, appear “shifted” towards problems of heath care, both in technology/scientific and social aspects.

A special emphasis should be given to the problem of words and NPs, which belong to a common-sense linguistic world model and, as a result, contribute little to understanding of the given subject domain, e.g. words like ‘use’ and ‘time’. However, by looking deeper into selected abstracts, one could see that even those commonsense words convey something essential to the contents of articles. Indeed, a scientific abstract is by its definition is a list of condensed and refined statements of the article. As a result, in a deep NLP of abstracts, an impact of 'parasitic' insignificant words is negligibly small.

4. Computation of pairwise similarity measures between NPs

A number of similarity measures, given both a certain statistics of term/word usage and a hierarchy of terms/words, are proposed in literature. Most known is Resnik similarity [14], which is defined as information content of the most specific taxonomy node, parental to both of given words, $c1$ and $c2$. The information content is defined as:

$$IC(x) = -\log\big(p(x)\big) \tag{4.1}$$

where x is a word and $p(x)$ is the probability of that word in a corpus of texts used.

Instead of computing a set of $\{p(x)\}$ from a substantial subset of Pubmed articles (an approach utilized in [4]), we just use here a publicly available Brown Corpus [5], which can be regarded as a good approximation to the actual purpose of the presented approach.

Another similarity metric, called Lin, is proposed by [15,16]:

$$\text{Lin}(c1, c2) = 2\times\text{ResnikSimilarity(c1,c2) / (IC(c1)+IC(c2))} \tag{4.2}$$

Just by looking onto its formula, one can see that, compared to Resnik Similarity, it smooths out impact of too rare or too frequent terms in the corpus, focusing on the actual relationship within a hierarchy.

The Brown Corpus enriches the WordNet hierarchy with values of information content for each term. As a result, one gets a better correlation with experts' judgement of similarities, than by using (purely linguistic) trees of synonyms of raw WordNet.

There are two more reasons to choose the Lin metric. First, it correlates well with a 'expert's intuition over semantic similarity of terms [16]. Second, the Brown corpus does not support the TF*IDF metric, which is a de facto standard in the field. Also, certain studies, e.g. [17] show that the Lin and similar metrics compete well with TF*IDF.

To apply that data to NPs, containing multiple words, one has to compute weight coefficients, by which the relative impact of *defining* words, compared to the *defined* one, decreases from right to left according to the language use of left definitions.

Suppose we have two NPs, a and b, which have m_a и m_b adjectives and n_a и n_b nouns correspondingly. The weight coefficient to be used in comparing words at positions i_a и i_b, counted from right to left, starting at 0, is defined as

$$C_{i_a i_b} = \frac{1}{1 + w^a_{i_a} + w^b_{i_b}} \tag{4.3}$$

where $w^a{}_{ia} = i_a$, $w^b{}_{ib} = i_b$. That is, for the defined words of NPs, which stand at rightmost positions, that coefficient equals 1. In comparing the defined word of one NPs to the first defining word of another it would equal 0.5, for two first defining words it would equal 0.33 and so on.

The resulting measure of similarity between two words of NPs to compare, is

$$S_{i_a i_b} = C_{i_a i_b} \text{Lin}\big(W_{i_a}, W_{i_b}\big) \tag{4.4}$$

That measure takes values from 0 to 1 even at $C_{i_a i_b} = 1$.

The final measure of similarity between NPs as wholes is defined iteratively. Starting from 0, we subsequently add one $S_{i_a i_b}$ after another, with the use of a formula of the sum of probabilities, which guarantees, that the result is always less or equal to 1:

$$S_1 + S_2 - S_1 S_2 \tag{4.5}$$

From now on, for the purpose of illustration example, we use ‘data mining’ as a query to Pubmed, which constrains a set of abstracts to work on (around 30000 abstracts in size). For some NPs which are found in results of that query, we got following similarity measures:

S(drug discovery, finding) = 0.96, where
S('drug.n.01', 'determination.n.01') = 0.0
S('discovery.n.01', 'determination.n.01') = 0.96

S(data mining method, data mining technique) = 0.93, where
S('data.n.01', 'data.n.01') = 0.2 (weight = 1 / (1 + 2 +2))
S('data.n.01', 'mining.n.01') = 0.01
S('data.n.01', 'technique.n.01') = 0.02
S('mining.n.01', 'data.n.01') = 0.01
S('mining.n.01', 'mining.n.01') = 0.333 (weight = 1 / (1 + 1 + 1))
S('mining.n.01', 'technique.n.01') = 0.08
S('method.n.01', 'data.n.01') = 0.03
S('method.n.01', 'mining.n.01') = 0.09
S('method.n.01', 'technique.n.01') = 0.838

A syntax construct of <word>.n.<number> is used in WordNet to denote a set of synonyms (a synset) for <word> at position <number> in the list of synsets for a certain vocabulary entry of that thesaurus. For example, ‘knowledge’ has ‘cognition.n.01’ as the main (the first) synset in WordNet. The ‘n’ symbol denotes a noun.

The second of two examples above shows how impact of coinciding defining words (‘data’ and ‘mining’) is decreased to a ‘reasonable’ level by means of weight coefficients. Otherwise, similarity measure for ‘data mining method’ and ‘data mining technique’ would be equal to 1, thus contradicting to any reasonable intuition. On the other hand, a commonality in left definitions ‘raised’ the similarity between the defined words ‘method’ and ‘technique’ from 0.83 to 0.93.

5. Further Processing of the Similarity Measures

First of all, let’s transform similarities to dissimilarities:

$$D_{ij} = 1 - S_{ij} \tag{5.1}$$

For a subset of NPs from the example above one gets (actual words/terms are of no essence):

Table 2.
Nominal Similarities

	kw1	kw2	kw3	...	kw7	kw8	kw9	kw10
kw1	0.0000	0.9374	0.9305	...	0.9985	0.8857	0.9975	0.9274
kw2	0.9374	0.0000	0.9421	...	0.9995	0.9076	0.9980	0.9416
kw3	0.9305	0.9421	0.0000	...	0.9995	0.8972	0.9980	0.9346
kw4	0.9383	0.9473	0.9433	...	0.9990	0.8340	0.9975	0.9419
kw5	0.8899	0.9102	0.9001	...	0.9985	0.5966	0.9980	0.8971
kw6	0.9152	0.9331	0.9247	...	0.9990	0.7791	0.9965	0.9222
kw7	0.9985	0.9995	0.9995	...	0.0000	0.9990	0.9985	0.9985
kw8	0.8857	0.9076	0.8972	...	0.9990	0.0000	0.9975	0.6409
kw9	0.9975	0.9980	0.9980	...	0.9985	0.9975	0.0000	0.9975
kw10	0.9274	0.9416	0.9346	...	0.9985	0.6409	0.9975	0.0000

One can see that nominal dissimilarities are distributed quite inhomogeneously. Most values tend to group around the higher ones. To remedy that we introduce an artificial dissimilarity, which takes values from 0 to 10 and is distributed normally with the mean of 50 and the standard deviation of 20. To map nominal values to ‘normalized’ ones, we match values which have same percentiles at both distributions. A similar approach is used in questionnaires like 16PF [18,19]. Therefore, the above example takes the form of:

Table 3.
Normalized dissimilarities

	kw1	kw2	kw3	kw4	kw5	kw6	kw7	kw8	kw9	kw10
kw1	0	48	45	50	31	40	74	30	62	44
kw2	48	0	53	55	39	46	99	37	66	51
kw3	45	53	0	54	36	43	99	35	66	47
kw4	50	55	54	0	23	0	83	28	62	52
kw5	31	39	36	23	0	20	74	9	66	33
kw6	40	46	43	0	20	0	83	25	56	41
kw7	74	99	99	83	74	83	0	83	74	74
kw8	30	37	35	28	9	25	83	0	62	16
kw9	62	66	66	62	66	56	74	62	0	62
kw10	44	51	47	52	33	41	74	16	62	0

That way, nominally close values are separated according to their actual statistical presence in the sample.

Resulting matrices of dissimilarities are then processed by a method of multidimensional scaling, which was developed for similar purposes, that is, to analyze qualitative (e.g. subjective) similarity judgments [4,12]. That is, every object (kw_i) is placed into a Euclidean space (of small to moderate number of dimensions) so that all distances in pairs of objects, corresponding to one and the same normalized dissimilarity, would be as equal as possible, and in the same time be greater than all distances which correspond to smaller dissimilarities.

A cluster analysis is then performed on top of resulting object vectors. A method of cluster analysis used, is described in literature [20], but quite rarely is met in standard libraries. That method results in quite well balanced cluster trees, which facilitates their expert interpretations.

The resulting binary tree is cross sectioned in a way, so that we have a more aggregated tree, consisting of a fixed number of levels, e.g. 6, where each of non-terminal nodes has a fixed maximum number of subclusters, e.g. 3. The terminal nodes may consist of arbitrary number of objects (NPs), which is defined solely by the distance in the obtained multidimensional space.

The two aggregation steps described, have a purpose to align the resulting structures as much as possible to ones, which could have been obtained via expert judgment.

A representative subtree, which we obtain by processing abstracts, resulting from the ‘data mining’ query to Pubmed, is displayed at figure 1.

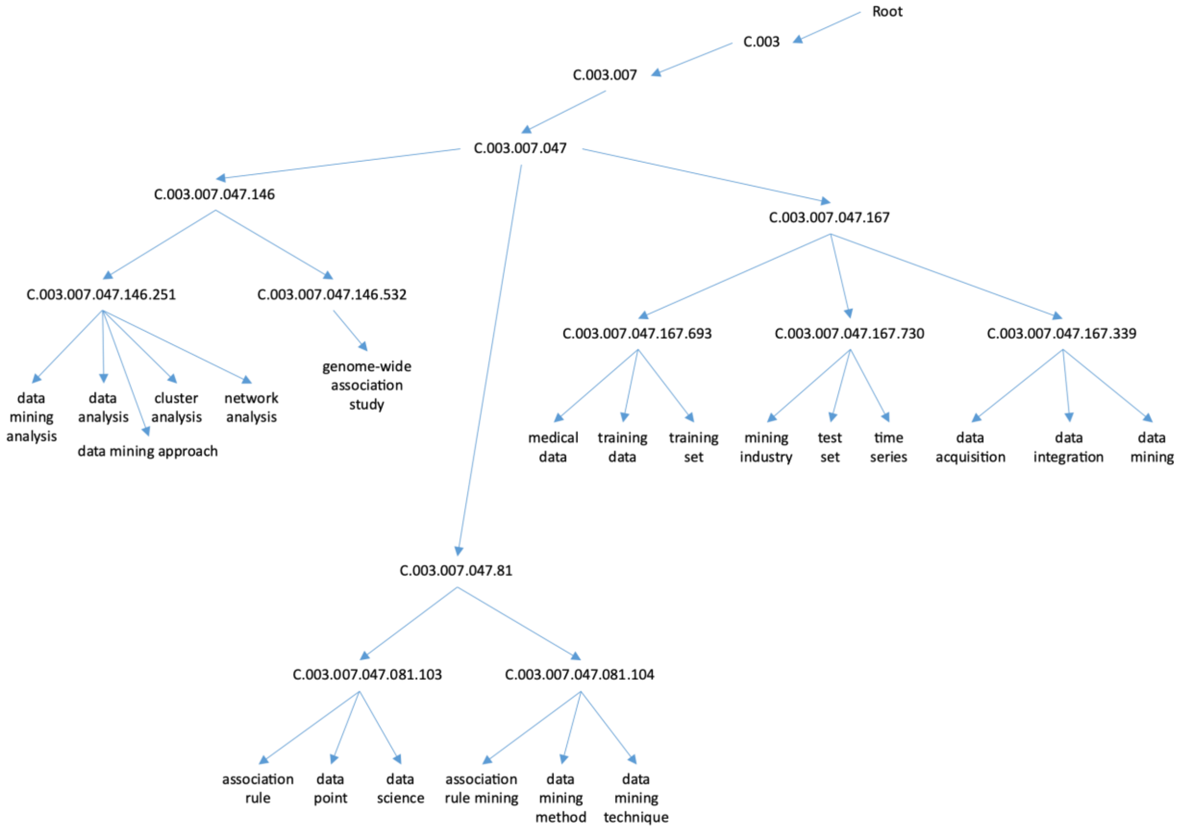


Figure 1: A fragment of the resulting hierarchy of NPs for the set of abstracts, matching the ‘data mining’ query.

That subtree looks quite compliant to an intuition of somebody, whose work is deeply or even slightly related to data mining.

However, that subtree is not free of false positives, the most striking of which is ‘mining industry’, cause by a polysemy of the word ‘mining’. Another one is less obvious (‘genome-wide association study’) and may require further research.

In any case, a human curation, accompanied by an incremental learning of correction samples, could significantly improve the quality of the hierarchy at question.

In general, one could see, that using NPs rather than individual words, despite ‘false positives’, brings useful enough semantic elements, which help building a meaningful thesaurus of a subject domain without using any apriori ontology, applicable to the chosen domain. The only apriori set of semantic data we used, is the Brown Corpus, which reflects a common sense linguistic world model. That specific corpus can be replaced, in subsequent research, by one, built from the texts, belonging to the chosen subject domain (may be a wider one for better statistics).

6. Extending MeSH with Noun Phrases, specific to the document sample

Each MeSH term has several attributes. A textual description of the term is one of them. For example:

Table 4

Selected MeSH Terms with descriptions

Label	Description
Logic	The science that investigates the **principles** governing correct or reliable **inference** and deals with the canons and criteria of **validity** in **thought** and demonstration. This system of reasoning is applicable to any branch of **knowledge** or study. (Random House Unabridged Dictionary, 2d ed & Sippl, Computer Dictionary, 4th ed)
Latent Class Analysis	A **statistical** algorithm used to analyze **clusters** of observed variables by constructing **categorical** unobserved or latent segment based on weighted **analysis** and the **average** probabilities. Such latent **classes** are used to infer variables whose relationships are not directly observed. In **biomedical** research, it is often used to categorize **data** that allows the determination of symptom **clusters**.

Given a set of NPs, extracted from a sample of abstracts, and a set of MeSH terms, found in annotations of those abstracts (that is, belonging to *that same* sample), we can walk through all pairwise matchs between NPs and the descriptions of those MeSH terms. In that case, it does not make sense to extract NPs from descriptions and apply weights to their left definitions, because descriptions are, in general, terse and short in length. We apply weights here only to left definitions of NPs, extracted from abstract texts, using Eqs. 4.3 and 4.4, and sum up all individual similarities between NPs and words of term descriptions. Resulting measure, that way, can exceed 1.

Words, highlighted in bold in the table 4, hint at matches, which led to extended hierarchies, shown at figures 2 and 3. In those, MeSH terms are shown in blue, whereas NPs are shown as bold red.

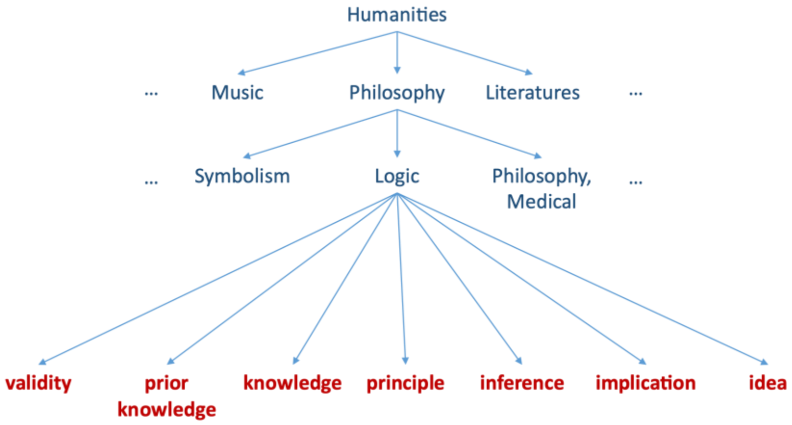
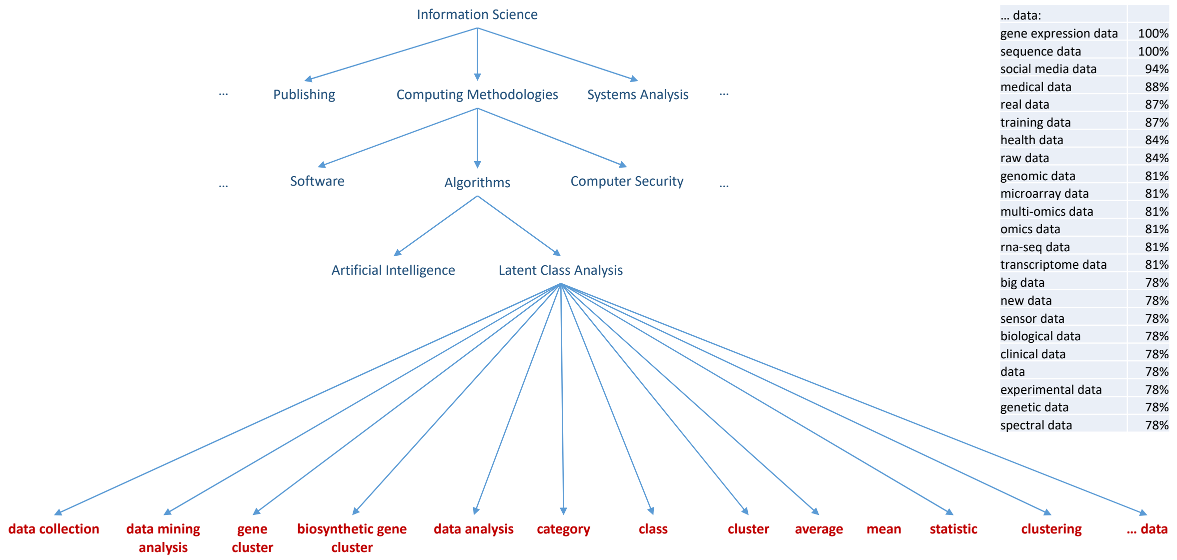


Figure 2: A subtree of MeSH, extended with NPs, extracted from a sample of abstracts



... data:	
gene expression data	100%
sequence data	100%
social media data	94%
medical data	88%
real data	87%
training data	87%
health data	84%
raw data	84%
genomic data	81%
microarray data	81%
multi-omics data	81%
omics data	81%
rna-seq data	81%
transcriptome data	81%
big data	78%
new data	78%
sensor data	78%
biological data	78%
clinical data	78%
data	78%
experimental data	78%
genetic data	78%
spectral data	78%

Figure 3. Another subtree of MeSH, extended with NPs, extracted from a sample of abstracts

A '…data' node at Figure 5.2 and a table above it represents matches of NPs, ending with 'data', to the 'Latent Class Analysis' MeSH term. The table is sorted by the similarity measure, normalized to a maximum, achieved at 'gene expression data' and 'sequence data'.

At the two examples above, especially at the second one, we see how a 'bibliographic', relatively shallow hierarchy of MeSH terms is extended with new terms (NPs), which tie that hierarchy to a specific, vocabulary-rich subject domain (life sciences and medicine in that case).

It must be noted, however, that a purely linguistic matching we use, can lead to a large number of false positive matches. E.g. 'biosynthetic gene cluster' has very little to do with data mining and data analysis, to which the 'Latent Class Analysis' belongs. Other false positives, as mentioned above, can result from even more misleading matches, which arise from polysemy or metonymy.

Nevertheless, the very ability to tie MeSH to a specific sample of Pubmed abstracts can significantly improve quality of the literature mining approach, proposed in [4].

7. Conclusion

The proposed approach to build a conceptual structure of a subject domain in the form of an ad hoc hierarchical thesaurus meets the original requirement: synthesis of a larger number of relatively meaningful and interrelated hierarchical term structures, which can be further dealt with as whole objects.

However, the approach, as it seems, has a significant potential for separate uses as a tool of knowledge engineering.

First, the associations it reveals between NPs of natural texts, well correlate both with intuition of builders of the Brown Corpus as well as one of experts who annotate Pubmed articles.

Second, a hierarchical structure like we discuss, may serve as a good initial approximation in development of a full-fledged ontology of the subject domain.

Also, a thesaurus like one built here, but comprising a large (more than 100000) number of terms, can serve as a better basis for literature mining based on multidimensional scaling and cluster analysis [4], than a 'librarian', by its purpose, MeSH thesaurus, used in the mentioned paper.

Further development of the discussed method, could, first of all, deal with the problem of false positives, e.g. a cluster of ('traditional method', 'conventional method', 'statistical method') or ('computational approach', 'traditional approach', 'new approach'), which obviously look quite loose to serve as a basis of a 'good' ontology. Those false positives could be eliminated by using a semi-automated training via any of the modern ML/NN methods of supervised learning.

Moveover, heuristics behind choices of data processing methods and definitions of metrics could be enriched with expanding a set of alternatives in conjunction with a ‘good’ meta-metric which would justify the final choice.

The current paper describes a first step in the directions outlined.

8. References

[1] Rogers, F. B. "Medical subject headings". Bull Med Libr Assoc. (1963): 114–116. Vol 51.

[2] Campos P.M.C., Reginato C.C., Almeida J.P.A., Barcellos M.P., de Almeida Falbo R., Souza V.E.S., Guizzardi G. “Finding reusable structured resources for the integration of environmental research data.” Environmental Modelling and Software (2020): 133.

[3] Voinov A.V. Ontology integration and elicitation of holistic knowledge (in Russian). Novosti Iskusstvennogo Intellecta. 2. 2005.

[4] Voinov A.V., Demikova N.S., Kobrinsky B.A. “Intellectual data analysis of medical data with the use of vocabulary scaling.” Twelfth National Conference with International participation KII-2010. (2010): 153-160.

[5] Francis, W. Nelson & Henry Kucera. BROWN CORPUS MANUAL: Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English for Use with Digital Computers. – http://icame.uib.no/brown/bcm.html. 1979.

[6] Alani H, Kim S, Millard DE et al. “Automatic ontology-based knowledge extraction and tailored biography generation from the web.” IEEE Intell Syst (2002): 18: 14–21.

[7] Makki J, Alquier A-M and Prince V. Ontology population via NLP techniques in risk management. Int J Hum Soc Sci 2009; 3:212–217.

[8] Faria C, Serra I and Girardi R. “A domain-independent process for automatic ontology population from text.” Sci Comput Program (2013): 95: 26–43.

[9] Hearst MA. Automatic acquisition of hyponyms from large text corpora. In: International conference on computational linguistics, Nantes, 23–28 August 1992, pp. 539–545. Stroudsburg, PA: ACL.

[10] Finkelstein-landau M and Morin E. “Extracting semantic relationships between terms: supervised vs. unsupervised methods.” International workshop on ontological engineering on the global information infrastructure, Dagstuhl Castle, Germany, (1999): 13 May, pp. 71–80.

[11] Yangarber R and Grishman R. “NYU: description of the Proteus/PET system as used for MUC-7 ST.” Message understanding conference, Fairfax, VA, (1998): 29 April–1 May.

[12] Chernigovskaya T.V., Gavrilova T.A., Voinov A.V., Strel’nikov K.N. Sensorymotor and cognitive laterality profile. Fiziologiia Cheloveka 31(2):24-33 (2005)

[13] Voinov A.V. Modeling experts’ intuition by methods of psychosemantics and inference with uncertainty (in Russian). Novosti Iskusstvennogo Intellecta. 2. p. 130. (1998)

[14] Resnik Ph. “Using information content to evaluate semantic similarity in a taxonomy.” Chris S. Mellish (ed.). Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95). 1: 448–453. (1995)

[15] Li Y., McLean D., Bandar Z.A., O’Shea J.D., Crockett K. “Sentence Similarity Based on Semantic Nets and Corpus Statistics.” IEEE Transactions on Knowledge and Data Engineering. August (2006): 1138-1150, vol. 18.

[16] Pedersen T. Information Content Measures of Semantic Similarity Perform Better Without Sense-Tagged Text. Department of Computer Science University of Minnesota, Duluth. tpederse@d.umn.edu http://wn-similarity.sourceforge.net . 2010.

[17] Michał Marcińczuk, Mateusz Gniewkowski, Tomasz Walkowiak, Marcin Będkowski, “Text Document Clustering: Wordnet vs. TF-IDF vs. Word Embeddings.” Proceedings of the 11th Global Wordnet Conference. UNISA. (2021): Pp. 207–214.

[18] Cattell R.B. Personality and Mood by Questionnaire. San Francisco, CA: Jossey-Bass. 1979.

[19] MMPI-2 Scales". University of Minnesota Press. Retrieved 24 April 2015.

[20] Jambu M. Classification Automatique pour l’Analyse des Donnees. Methodes et algorithmes. Dunod. 1978.