FOLKSONOMY - SUPPLEMENTING RICHE EXPERT BASED TAXONOMY BY TERMS FROM ONLINE DOCUMENTS (Pilot Study)

Aleš Bourek, Mikhail Alexandrov, Roque Lopez

***Abstract**: RICHE (Research Inventory of Child Health in Europe) is a platform developed and funded under the Health domain of 7th European Framework Program. The platform search engine is expected to use the multilingual taxonomy of terms for processing and classifying large volumes of documents of the RICHE repository. So far the experts participating in this project have produced the initial version of expert based taxonomy of terms relating to child health (based on existing taxonomies). In the paper we propose a simple man-machine technique for continuous support and development of the existing term list, which consists of three steps: 1) construction of various keyword lists extracted from a topic oriented document set using various levels of word specificity 2) selection of the most useful keyword lists using subjective criteria as a precision of selection and a number of new words 3) manual selection of new terms. Experimental material was represented by documents uploaded from three organizations active in child health improvement policies: World Bank, World Health Organization (WHO), and DG SANCO of European Commission (EC). The selection was performed in order to assess terms used in these documents that may be absent in the RICHE taxonomy. Presented work should be considered as a pilot (feasibility) study. The objective of the RICHE platform is to identify gaps in European child health research, so extensive mapping exercise has been started. Classification of identified studies is essential and cannot be based only on traditional terms of existing taxonomies. Emergent terms (such as for example "cyberbullying") need to be identified and included into existing taxonomies. In our future work we focus on developing techniques related to multilevel and multiword term selection*

***Keywords**: Child health, natural language processing, taxonomy, term selection*

***ACM Classification Keywords**: I.2.7 Natural Language Processing*

Introduction

1.1 RICHE project and its current taxonomy

The European Commission (EC) and other funding agencies make large investments in child health research. The health of out children is satisfactory, but there are serious concerns, for example obesity, mental health, alcohol abuse, and sexuality. The objective of EC efforts is to establish a sustainable network for researchers, funders, policy makers, advocates and young people in Europe to support collaboration in developing the future of child health research. In the RICHE project we are producing an inventory of research and reports on gaps in research and on roadmaps for future research [RICHE, http].

RICHE platform includes a search engine for efficient retrieval of information included in the platform and related to child health protection and child health and healthcare quality improvement. The effective function of this engine needs well prepared lexical resources based on specific terms reflecting the topic under consideration, for example, appropriate taxonomies [Taxonomy, http]. It was found that "in some cases, it was essential to look outside traditional health and social care search engines in order to fully understand and conduct a systematic review on subjects that are relevant and pertinent to public health, such as justice and police databases. As far as the taxonomy structures were concerned, the existing classification identified limited indexing used in some databases as a potential problem - "Where free text words are relied upon, variations in terminology used by different disciplines can create barriers and limit the value of the material retrieved in a search of the database/document repository" [RICHE, http].

To address this issue child health experts, mainly from the area of public health research, participating in RICHE project have prepared the initial version of taxonomy consisting of one-word and multiword terms distributed on 6 sub-topics (main axis of child health determinants): 1).Demographics 2).Population group 3).Agents, influences and settings 4).Health, disability, health issues and determinants 5).Language 6).Type of study. Table 1 shows some terms from the sub-topic 'Demography' as a part of the mentioned taxonomy

Table 1. Example of terms of sub-topic 'Demography' taken from RICHE taxonomy

Terms	Synonyms
Indeterminate / anomalous	unknown, unstipulated, uncertain, not determined at birth
Stillborn children	died before birth, died in utero, born dead
Genetic studies	heredity, inherited, chromosomal, inborn, genomic

It is obvious, that the RICHE taxonomy contains multiword terms located on two levels corresponding to given 6 sub-topics. Totally the current RICHE taxonomy contains 822 different one-word terms in stem-form. Currently RICHE experts continue to improve the taxonomy in two directions: modifying the term list and constructing a more detailed hierarchy

2.2 Problem settings

Goal of our contribution is to identify and add new terms to the existing "expert taxonomy" using appropriate NLP tools. On the given stage of our work we introduce two limitations:

- we deal with one-word terms and one-level term distribution
- we process limited number of documents ,

The first limitation is introduced by the fact that multiword and multilevel term list construction requires the use of sophisticated methods but as a feasibility study we decided to address the problem using as simple as possible tools. The second limitation is defined by our approach to analyze (when necessary) individual documents but not to work with descriptive statistics.

In the paper we propose a simple methodology for augmenting the existing term list constructed by RICHE experts, and to test this technique experimentally. The technique consists of 3 steps:

- construction of various keyword lists extracted from the above mentioned document set using various levels of word specificity
- selection of the most useful keyword lists using subjective criteria for the accuracy of selection and a number of new words to be analyzed
- manual selection of new terms by an expert

To demonstrate possibilities of the proposed methodology we performed experiments with different subsets of documents. The source of information for our experiments were 60 documents (7 Mb in plain textual format) downloaded from online resources of World Bank [World Bank, http], World Health Organization [WHO, http], and European Commission [EC, http]. These organizations belong to main policy players in the area of child health in Europe.

The auxiliary problem we studied was the dependence of results on a concrete document set, which was used as a source of new terms. For this we considered various subsets of a given document corpus and compared their lexical resources from the point of view of our main goal - improvement of existing RICHE term list.

1.3 Related work

Indicators of child health were introduced and studied in many projects, for example, [Rigby, 2002; Rigby, 2003]. These indicators need information reflecting current state of child health and RICHE platform is supposed to provide this information.

There are several works related with term selection focused on medical applications [Madden, 2007; Armstrong, 2009]. But our task is different: to supplement the existing term list by new terms from independent sources such as the Internet or the domain of "gray literature".

The key position in problem solution consists in constructing various keyword lists for consideration for further detailed analysis by a child health expert(s). The general approaches and algorithms for term selection are well presented in many publications, for example in the well-known monograph [Baeza-Yates, 1999]. An interesting approach to multilevel term selection is described in [Makagonov, 2005]. It is recognized that word collocations have a large informative and distinctive power. Just these collocations form so-called multiword terms [Yagunova, 2010]. But all these techniques are not simple. They often need complimentary information about word distribution in a corpus, correlation between words, etc. In this paper we deal with the simplest case: one-word and one-level term selection.

We apply the criterion of word specificity for extraction of keywords (candidates to be included in term list) from a given document set. This criterion was successfully used for constructing domain oriented vocabularies [Makagonov, 2000]. Recently free-share LexisTerm program was developed [Lopez, 2011] where both a traditional corpus based option and the new document based option are used [Lopez, 2011]. We use both of these options in our work.

In section 2 we describe the proposed methodology. In section 3 we demonstrate the results of experiments. Section 3 includes conclusions.

2. Methodology of term selection

2.1 General description

We use word 'keyword' instead 'term' on the stages of constructing initial keyword lists and selecting the best lists for further manual analysis. Here the selected keywords are only the "candidates to be" terms if an expert will select them.

As mentioned in introduction the proposed methodology consists in three stages:

1) Constructing several keyword lists on the basis of criterion of word specificity. We use the criterion of word specificity because the topic under consideration is not broad enough and we expect to obtain more or less useful keyword lists. But we do not know in advance what level of specificity and what option of selection will prove to be the most relevant to the existing expert list. For this reason we have to generate several keyword lists.
2) Selecting the most useful keyword lists. Here we compare each keyword list constructed on the previous stage with the expert list using indicator of precision and the number of keywords not included in the expert list (external keywords). We use indicator of precision to be more confident that not-common keywords are relevant to the expert list. But in general high precision refers to the case of very short keyword lists with very high level of word specificity. Such short lists can be un-useful. In this case an

expert must evaluate the number of non-common words and makes a decision whether the concrete keyword list is useful or not.

3) Extracting terms from the keyword lists selected on the previous stage. This is performed manually by an expert.

In our study we used stems instead of original word forms.

The auxiliary research concerned studying the dependence of results on concrete document sets. Here we performed two simple experiments with different document sets

- Comparison of keyword lists extracted from the half and from all documents with the expert list on the basis of indicators of precision and recall
- Comparison of keyword lists between themselves (without taking into account the expert list) constructed for a subset of 15% documents and for a different 15% subset of documents. The same procedure was implemented for 30% document subset and other 30% document subset, and finally for 50% documents and other 50% document subset. We use here only indicator of precision with respect to each document subset from the pair.

In these experiments we used the same fixed parameters for keyword extraction: level of keyword specificity and option of keyword selection.

2.2 Constructing keyword lists by the criterion of word specificity

To construct keyword lists we used the LexisTerm [Lopez, 2011] program. Following are some necessary definitions:

Definition 1. The general lexis is a frequency word list based on a given corpus of texts

The given corpus means here any standard document set reflecting the lexical richness of a given language. Generally such a corpus contains in a certain proportion the documents taken from newspapers, scientific publications related with various domains, novels and stories. For example, it could be the British National corpus.

Definition 2. The level of specificity of a given word **w** in a given document corpus C is a number $K \geq 1$, which shows how much its frequency in the document corpus $f_C(\boldsymbol{w})$ exceeds its frequency in the general lexis $f_L(\boldsymbol{w})$:

$$K = f_C(\boldsymbol{w}) / f_L(\boldsymbol{w})$$

Definition 3. The level of specificity of a given word **w** in a given document D is a number $K \geq 1$, which shows how much its frequency in the document $f_D(\mathbf{w})$ exceeds its frequency in the general lexis $f_L(\boldsymbol{w})$:

$$K = f_D(\boldsymbol{w}) / f_L(\boldsymbol{w})$$

Our research was done using keywords and terms presented in stem form. For this we had to transform both documents and general lexis based on British National corpus to their stem using the well-known Porter stemmer [Porter, 1980]

2.3 Measures for comparison of word lists

To select the most preferable lists of keywords we used two variables: precision of keyword selection and the number of keywords in the list not included in the expert list. Following is a description of these variables:

Let N_L is a number of terms in the expert list, N_W is a number of keywords in our list, N_{LW} is a number of common words in both lists. In this case a precision is calculated according the formula:

$$P = N_{LW} / N_W$$

That is the precision, it is a share of terms from the expert list in our keyword list. With the designation intruduced above the number of new keywords in our list N is calculated by the formulae

$$N = N_W - N_{LW}$$

Additionally an expert can take into account the other two indicators of quality used in Information Retrieval: recall and so-called F-measure. They are calculated according the following formulae:

$$R = N_{LW} / N_L$$

$$F=2(PR) / (P+R)$$

In the auxiliary experiments we needed to compare two keyword lists. Let $N1_W$, $N2_W$, $N12_W$ be the number of keywords in the 1-st list, 2-nd list and the common keywords respectively. In this case one considers the precision with respect to each list and the average precision. They are calculated according the formulae:

$$P_1 = N12_W / N1_W$$

$$P_2 = N12_W / N2_W$$

$$P_{12} = (P_1+P_2) / 2$$

All these indicators are used in the experiments described in the next section

Experiments

3.1 Selection of preferable keyword lists

In this experiment we compared keywords selected from our full document set consisting of 60 documents with the full expert list. We constructed keyword list under different levels of word specificity (*k*=1,5,10,20,50,100) and different options (C and D). The results are presented in Table 2. The designations in this table are described in the previous section. Figure 1 shows a graphical view of Table 2 for the precision

Table 2. Characteristics of different keyword lists, comparison with the complete expert list

	C, k=1	***C, k=5***	***C, k=10***	***D, k=10***	***D, k=20***	***D, k=50***	***D, k=10***
Words	1473	231	86	1807	1193	512	219
P	0.238	0.442	0.558	0.200	0.226	0.252	0.324
R	0.426	0.124	0.058	0.440	0.328	0.157	0.086
F	0.305	0.194	0.106	0.275	0.268	0.193	0.136
N_{LW}	350	102	48	362	270	129	71
N	1123	129	38	1445	923	383	148

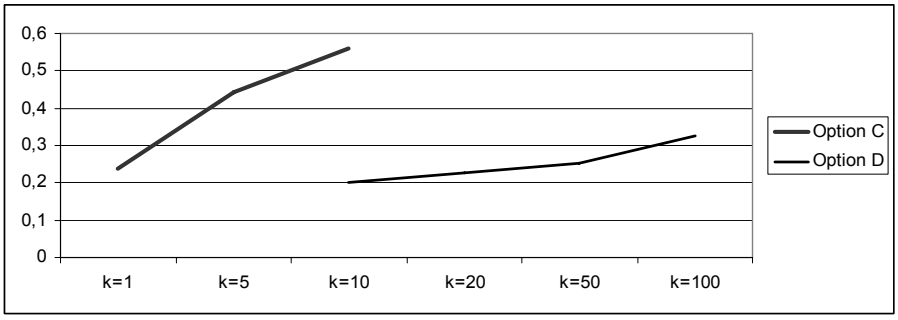

Fig.1 Graphical view of the Table 2 for the values of precision

One of the authors, an MD also engaged in medical informatics selected the two most useful/preferable lists with the parameters: option C, *k*=10 and option D, *k*=100. With these parameters we obtained the highest precision

in the framework of given mode, and from the other hand the number of new non-common keywords is suitable for manual evaluation.

3.2 Contribution of sub-topics to keyword lists

In this experiment we compared our keyword lists with the terms of experts related with each category. The results are presented in the Table 3. Here we consider option C with the parameters *k*=5, 10. Figure 2 provides a graphed version of Table 2.

Table 3. Characteristics of different keyword lists, comparison with each category of the expert list

	Category 1	***Category 2***	***Category 3***	***Category 4***	***Category 5***	***Category 6***
Option C,k=5	0.030	0.065	0.121	0.247	0.061	0.078
Option C,k=10	0.047	0.093	0.128	0.291	0.116	0.105

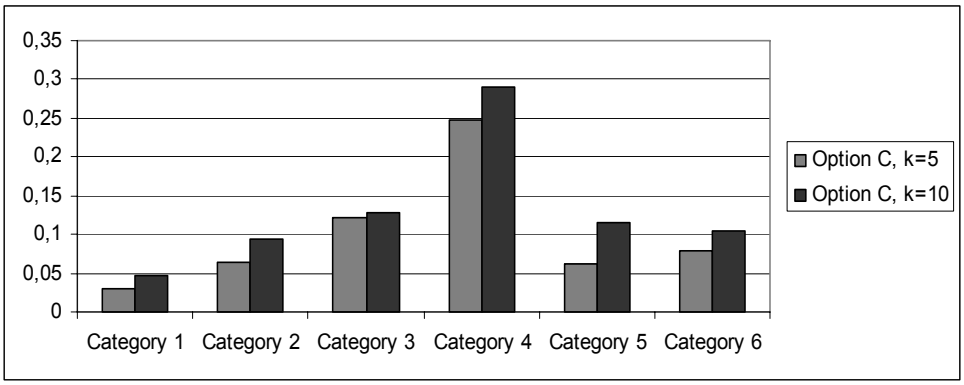

Fig.2 Graphed version of Table 2

It is easy to see that Category 4 (RICHE taxonomy axis) 'Health Disability') is represented the best way in our keyword lists. Such result is not surprising, since this axis is the most comprehensive in included terms.

3.3 Lexical resources of subsets

In the first experiment we compare keyword lists extracted from the half and from all documents with the expert list on the basis of indicators of precision, recall and F-measure. The results are presented in the Table 4. In the second experiment we compared keyword lists for different pairs of document subsets using precisions with respect to each subset and the averaged precision. The results are presented in the Table 5. The designations of these tables are described in the previous section. In all experiments we used option C with the level of specificity *k*=1.

Table 4. Comparison of lexical resources of different document number with the expert list

	50% documents	***100% documents***
Words	1019	1473
P	0.277	0.238
R	0.343	0.426
F	0.306	0.305

Table 5. Cross-comparison of lexical resources of different document subsets

	10 +10	***20+20***	***30+30***
Words	959/1171	1065/1234	1019/1333
P_1	0.766	0.783	0.838
P_2	0.628	0.676	0.656
P_{12}	0.697	0.729	0.747

Data of table 4 shows a naturally tendency: the more documents we consider, the more different words they have the less precision is. Table 5 demonstrates the relative stability of the results: any equal subsets of documents (having in view the number of documents) have close averaged precision. These circumstances inform about good quality of selected document corpus.

3.4 Term selection

As we have already mentioned in section 3.1 the preferable keyword lists for term selection prove to be those with the parameters: option C, *k*=10 and option D, *k*=100 – offering enough potential "candidate" terms for inclusion into the RICHE "expert taxonomy" but at the same time producing only a small number of false identified terms.

A quick "human" selection of potential terms (word stems) for classification of child health research documents identified by the machine-learning methodology presented from a volume of text relating to child health from websites of WHO, DG SANCO (European Commission) and World Bank (policy related documents on child health) was performed.

Following list shows some, not all, "candidate" term stems (word stems NOT included in the original RICHE child health experts version of the taxonomy) were identified:

clinic *(possible candidate for classification of "clinical study, type of clinics")*
HIV *(not included and exists only in a synonym classification option as "AIDS")*
implement *(possible candidate for classification of "implementation research")*
monitor *(possible candidate for classification of "monitoring research")*
overview *(possible candidate for classification of "overview materials")*
agenda *(possible candidate for classification of "agenda setting research")*
cognit *(possible candidate for classification of "cognitive related research")*
complex *(possible candidate for classification of "complexity research")*
indicator *(possible candidate for classification of "indicator related research")*
consensu *(possible candidate for classification of "consensus based documents")*
emerg *(possible candidate for classification of "emergent issues related research")*
fat *(not even the synonym obes (obesity) is included as a term in the RICHE expert taxonomy)*
framework *(possible candidate for classification of "framework setting research")*
guideline *(possible candidate for classification of "clinical guidelines related research")*
Mediterranean *(often used geographical term, NOT included in the Language/Geography axis)*
pregnan *(possible candidate for classification of "pregnancy related research")*
priorit *(possible candidate for classification of "prioritization research, priority setting research")*
protocol *(possible candidate for classification of "protocol setting research")*
questionnaire *(possible candidate for classification of "questionnaire/survey related research")*
satisfact *(possible candidate for classification of "health service satisfaction research")*
vitamin *(possible candidate for classification of "vitamin related research")*
facilit */is included as a term in the RICHE expert taxonomy in the form of "Health care facility" BUT NOT*

for example as "facilitation study"/
***feed** /is included as a term in the RICHE expert taxonomy BUT only in the form of " breastfeeding"/*
***global** /is included as a term in the RICHE expert taxonomy BUT only in the form of "Global change and health (WHO-Europe)" BUT NOT as "globalization related research"/*

On the other hand, four terms identified as "missing" by the machine-learning based methodology were already included in the original RICHE "expert taxonomy":

***analys** /is included as a term in the RICHE expert taxonomy/*
***demograph** /is included as a term in the RICHE expert taxonomy/*
***expenditur** /is included as a term in the RICHE expert taxonomy/*
***outcom** /is included as a term in the RICHE expert taxonomy/*

Conclusion

We elaborated on and proposed the simplest way for supporting term list development experts of the RICHE project. Our methodology is based on criterion of specificity for keyword selection and characteristics of precision for keyword list selection having in view the possibilities of subsequent manual work of an expert. The results of our experiments may prove useful in evaluating how criterion parameters affect the list of selected terms.

Term stems expertly identified as possible classification term "candidates" have been correlated with terms of the RICHE_expert_taxonomy_ver_January_2011. The four term stems followed by the remark "/is included as a term in the RICHE expert taxonomy/" represent false identified terms by means of our simple machine learning based approach. With the exception of these four terms all above listed stems have a potential for classifying child health related research documents of the RICHE repository, as commented in the brackets following the respective term. All of the "candidate" terms will be presented to the RICHE consortium group for expert evaluation and for inclusion of terms the experts will find consensus on into the most appropriate axis of the RICHE project taxonomy. Based on the presented small scale preliminary analysis of 60 documents, we demonstrate that the methodology has the strength and potential to identify terms possibly missed by the expert community, especially when a corpus of documents produced by experts focused on a different area of child health (policy issues rather than public health child research – which was the dominant area of expertise of the majority of RICHE project collaborators) is used. We conclude that even basic machine-aided document evaluation is a tool for consideration when addressing the issue of possible human bias of the taxonomy defining expert community.

Bibliography

[Armstrong, 2009] Armstrong, R., Doyle, J., Waters, E. Cochrane Public Health Review Group Update: incorporating research generated outside of the health sector. Journ. of Public Health. Vol. 31, No. 1, pp. 187-189, 2009 (available at http:// jpubhealth.oxfordjournals.org/cgi/reprint/fdn116v1.pdf)

[Baeza-Yates, 1999] Baeza-Yates, R., Ribero-Neto, B. Modern Information Retrieval. Addison Wesley, 1999.

[EC, http] http:// ec.europa.eu

Lopez, R., Alexandrov, M., Barreda, D., Tejada, J. LexisTerm – the program for term selection by the criterion of specificity (this Proceedings)

[Madden, 2007] Madden, R., Sykes, C., Usten,T. World Health Organization Family of International Classifications, definitions, scope and purpose. 2007 (available at http://www.who.int/classifications/en/FamilyDocument2007.pdf)

[Makagonov, 2000] Makagonov, P., Alexandrov, M., Sboychakov, K. A toolkit for development of the domain-oriented dictionaries for structuring document flows. In: Data Analysis, Classification, and Related Methods, Studies in classification, data analysis, and knowledge organization, Springer-Verlag, pp. 83–88, 2000

[Makagonov, 2005] Makagonov, P., Figueroa, A., R., Sboychakov, K., Gelbukh, A. Learning a domain ontology from hierarchically structured texts. In: Proc. of Workshop "Learning and Extending Lexical Ontologies by using Machine Learning Methods" at 22-nd Intern. Conf. on Machine Learning (ICML 2005), Bonn, Germany, 2005.

[Porter, 1980] Porter, M. An algorithm for suffix stripping. Program, 14, pp. 130–137, 1980.

[RICHE, http] RICHE: http://childhealthresearch.eu

[Rigby, 2002] Rigby, M., Kohler, L. (edit.) Child health indicators of life and development (Child): report to the European Commission, 2002 (available at http: //www.europa.eu.int/comm/health/ph/)

[Rigby, 2003] Rigby, M., Kohler, L., Blair, M, Metchler, R. A priority for a caring society. Europian Journ. on Public Health, vol. 13, pp. 38-46, 2003

[Taxonomy, http] Taxonomy: http: //www.taxonomywarehouse.com/resultsbycat_include.asp?vCatUID=21&catcode=040100

[WHO, http] World Health Organization: http:// www.who.int

[World Bank, http] World Bank: http:// www.worldbank.org

[Yagunova, 2010] Yagunova, E., Pivovarova, L., The Nature of collocations in the Russian language. The Experience of Automatic Extraction and Classification of the Material of News Texts // Automatic Documentation and Mathematical Linguistics, 2010, Vol. 44, No. 3, pp. 164–175. © Allerton Press, Inc., 2010.

Authors' Information

	***Ales Bourek*** *– Senior lecturer, Masaryk University, Brno, Czech Republic; Head of Center for Healthcare Quality, Masaryk University. Kamenice 126/3, 62500 Brno, CZ. e-mail: bourek@med.muni.cz* *Major fields of interest: reproductive medicine – gynecology, health informatics, healthcare quality improvement, health systems*
	***Mikhail Alexandrov*** *– Professor, Academy of national economy and civil service under the President of Russia; Prosp. Vernadskogo 82, bld. 1, Moscow, 119571, Russia; fLexSem Research Group, Autonomous University of Barcelona, 08193 Bellaterra (Barcelona), Spain;* *e-mail: MAlexandrov@mail.ru* *Major fields of scientific research: data mining, text mining, mathematical modeling*
	***Roque López*** *– Student of System Engeenering at San Agustin National University, calle Santa Catalina Nº 117 Arequipa, Peru; e-mail: rlopezc27@gmai.lcom* *Major fields of scientific research: natural language processing, text mining, social network analysis*