Hybrid approach to knowledge extraction: textual analysis and evaluations of experts

Boris A. Kobrinskii
Institute of Modern Information Technologies in Medicine
Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
Moscow, Russia
kba_05@mail.ru

Nikolay A. Blagosklonov
Institute of Modern Information Technologies in Medicine
Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
Moscow, Russia
nblagosklonov@gmail.com

***Abstract*—The problem of extracting knowledge from medical experts for rare diseases causes difficulties due to the need to involve additional information about the clinical manifestations in this pathology. The report will consider a complex two-stage system of knowledge extraction. First, from literary sources using textological analysis. Then an expert evaluation of the information extracted from the literature with the addition of certainty factors.**

***Index Terms*—knowledge extraction, textological analysis, textological card, certainty factor, cognitive science world model, expert world model**

I. Introduction

Extraction of knowledge for intelligent medical systems supporting the adoption of diagnostic solutions in hereditary diseases is always characterized by considerable difficulties. This is due to the fuzzy of the changes observed in this pathology and the great similarity of a number of diseases. The complexity of differential diagnostics of orphan (rare) metabolic diseases is determined by genetic heterogeneity, clinical polymorphism (multivariate character space), constant progression (ie, continuous development) of diseases and absence of pronounced clinical signs at an early age. At the same time, early and accurate diagnosis is crucial for the timely prescription of pathogenetic therapy. At present, special attention is drawn to such a subject area as lysosomal storage diseases, in particular mucopolysaccharidosis, sphingolipidosis (GM1- and GM2-gangliosidoses, Gaucher disease, galactosialidosis, Farber granulomatosis, leukodystrophy, Niemann-Pick disease, etc.), mucolipidosis (I cell disease, etc.), glycoproteinoses (fucosidosis, sialidosis, mannosidosis, Pompe disease, Danone's disease, etc.) [1].

Accordingly, the purpose of this study is to extract knowledge about the occurrence and diagnostic significance of clinical signs of metabolic hereditary diseases for the subsequent constraction of an expert system for supporting the diagnostic process in different age periods.

The proposed two-stage procedure involves the knowledge extraction in consecutive order. At the first stage, a textological analysis of special literary sources (articles, monographs, etc.) is carried out to identify the signs (symptoms) characterizing the clinical manifestations of diseases in different age groups using a linguistic scale for assessing the degree of expression of qualitative features, at the second stage communicative methods of work with experts. To date, expert diagnostic systems for genetic diseases have been built on the basis of a dialogue between a cognitive scientist and an expert-physician and did not include a preliminary textological analysis of literary sources. It should also be noted that, unlike [2], the search for knowledge about the manifestations of diseases is carried out in manuals, monographs, articles, but not as a result of analysis of clinical guidelines and technological maps that determine the necessary patient examinations and the procedure for the actions in the process of diagnosis and treatment.

II. Formation of textological cards

To fix the results of textological analysis, a configuration of data representation in a textological card was developed (Table 1). For each nosological form of the disease, "textological cards" were formed, including a list of signs with their degree of severity and frequency of occurrence in different age periods. As an example, Table 2 presents a textological card for one of the lysosomal orphan diseases GM1-gangliosidosis.

TABLE I
Format of data representation in a textological card

Symptom	Intensity	Age group	Frequency of manifestation	Source (author)

The extraction of cognitive knowledge from texts is based on the identification of certain semantic fragments [3], in this case the signs and symptoms of hereditary diseases occurring at a certain age in the disease under consideration.

A textological card allows one to present knowledge in any structured document from any number of special (professional) publications with the indication of their authors. The authors' names also provide additional information for experts, since they allow us to indirectly assess the reliability of the fragments of knowledge extracted from the text in each area of medicine (neurology, orthopedics, cardiology, etc.). This is of great importance for the polysystemic diseases under consideration.

TABLE II
TEXTOLOGICAL CARD OF GM1 GANGLIOSIDOSIS

Symptom	Intensity	Age group[1]	Frequency of manifestation	Source (author)[2]
low height	strongly	1	frequently	1
coarse facial features	weakly	1	average	1, 2, 3, 4, 5
	strongly	2	frequently	
macroglossia	strongly	1	frequently	2
	strongly	2	frequently	
multiple mucopoly saccharidose-like dysostosis	weakly	1	average	2, 4, 5
	strongly	2	frequently	
	weakly	3	frequently	
stiffness of the joints	moderate	2	frequently	1
thickening of wrist joints	-	2	-	1
contractures of the elbow and knee joints	-	-	-	1
deformation of the hand like a "clawed foot"	-	3	-	1
kyphosis	strongly	2	frequently	1
spacity of muscles	strongly	3	-	1
	weakly	4	rarely	2, 6, 7
	weakly	5	rarely	
hypotonicity	strongly	1	frequently	1, 3, 5
	strongly	2	frequently	
ataxia	strongly	3	frequently	6
	weakly	4	rarely	2, 7, 8
	weakly	5	rarely	
convulsions	strongly	1	frequently	1, 2, 5, 6
	strongly	2	frequently	
	strongly	3	frequently	
backward mental development	strongly	1	frequently	1, 2, 3, 5
	strongly	2	frequently	
	strongly	3	frequently	6
neurogical disorders	strongly	2	-	1
	strongly	3	frequently	2, 6
	strongly	4	frequently	6, 7, 8, 9
	strongly	5	frequently	
cherry-red spots on retina	-	1	average	2, 3, 5, 6
	-	2	average	
	-	3	rarely	
corneal opacity	strongly	4	frequently	2
	strongly	5	frequently	
blindness	strongly	1	-	6
	strongly	2	-	
swelling	strongly	1	frequently	2
	strongly	2	frequently	
hepatomegaly	moderate	1	frequently	1, 2, 3 4, 5, 6
	moderate	2	frequently	
splenomegaly	moderate	1	frequently	1, 2, 3 4, 5, 6
	moderate	2	frequently	
cardiomyopathy	-	1	frequently	4
	-	2	frequently	

[1] 1 - at birth, 2 - 0-1 years, 3 - 1-3 years, 4 - 4-6 years, 5 - over 6 years

To assess the frequency of manifestation of features, a discrete (quantized) linguistic scale was developed that includes 5 levels [4]:

- 5 - very frequently (in 90 - 100%),
- 4 - frequently (in 70 - 80%),
- 3 - average (in 50%),
- 2 - rarely (in 20 - 30%),
- 1 - very rarely (in 10%),

In specific textological cards only the required scale levels are used. In the example of such levels there are only three (frequently, average, rarely).

III. WORLD MODELS OF THE COGNITIVE SCIENTIST AND EXPERT

B.M. Velichkovsky and M.S. Kapitsa [5] consider that the text in natural language is only a conductor of meaning, and the author's intention and knowledge lie in the secondary (semantic) structure of the text that is tuned over the natural text. Earlier it was noted that understanding is the formation of the "second text", that is, the semantic or conceptual structure [6]. Extraction of knowledge from texts is considered precisely as a task of understanding and highlighting the meaning of the text [7]. In the terminology of artificial intelligence is an attempt to recreate the semantic structure in the process of model formation. That is, it is the first step in structuring knowledge.

[2]The following sources are written in Russian and English languages were used in the preparation of textological card of GM1 gangliosidosis:

1. K. Jones, M. Jones, M. del Campo. Smith's Recognizable Patterns of Human Malformation. 7th Edition, Saunders, 2013, 1016 p.
2. D.L. Kasper, A.S. Fauci, S.L. Hauser, D.L. Longo, J.L. Jameson, J. Loscalzo. Harrison's Principles of Internal Medicine (19th ed.), McGraw-Hill Professional, 2015, 3000 p.
3. R.E. Berman. *Pediatriya. Rukovodstvo. V 8 knigakh. Bolezni ploda i novorozhdennogo, vrozhdennye narusheniya obmena veshchestv* [*Pediatrics. Guide. In 8 books. Diseases of the fetus and newborn, congenital metabolic disorders*], Moscow, Meditsina, 1991. 528 p.
4. A. Hinek, S. Zhang, A.C. Smith, J.W. Callahan. Impaired Elastic-Fiber Assembly by Fibroblasts from Patients with Either Morquio B Disease or Infantile GM1-Gangliosidosis Is Linked to Deficiency in the 67-kD Spliced Variant of b Galactosidase. American Journal of Human Genetics, 2000, vol. 67, no 1, pp. 23-36.
5. S.V. Serkov , E.Yu. Zaharova, G.N. Levitskij, I.N. Pronin. MRT pri pozdnei infantil'noi (yuvenil'noi) forme GM1-gangliozidoza. Nablyudenie iz praktiki [MRI of the Late Juvenile Form of GM1-gangliosidosis (A Case Report)]. *Meditsinskaya vizualizatsiya* [*Medical visualization*], 2006, no 1, pp. 123-127.
6. J. Campdelacreu, E. Muoz, B. Gmez, T. Pujol, A. Chabs, E. Tolosa. Generalised dystonia with an abnormal magnetic resonance imaging signal in the basal ganglia: A case of adult-onset GM1 gangliosidosis. Movement Disorders, 2002, vol. 17, no 5, pp. 1095-1097.
7. M. Hirayama, Y. Kitagawab, S. Yamamotob, A. Tokudaa, T. Mutoha, T. Hamano, T.Aita, M. Kuriyama. GM1 gangliosidosis type 3 with severe jaw-closing impairment. Journal of the Neurological Sciences, 1997, vol. 152, no 1, pp. 99-101.
8. U. Muthane, Y. Chickabasaviah, C. Kaneski, S.K. Shankar, G. Narayanappa, R. Christopher, S.S. Govindappa. Clinical features of adult GM1 gangliosidosis: Report of three Indian patients and review of 40 cases. Movement Disorders, 2004, vol. 19, no 11, pp. 1334-1341.
9. E. Roze, E. Paschke, N. Lopez, T. Eck, K. Yoshida, A. Maurel-Ollivier, D. Doummar, C. Caillaud, D. Galanaud, T. Billette de Villemeur, M. Vidailhet, A. Roubergue. Dystonia and Parkinsonism in GM1 type 3 gangliosidosis. Movement Disorders, 2005, vol. 20, no 10, pp. 1366-1369.

It should be specially noted that the semantic structures that the cognitive scientist distinguishes from the text include not only the author's world model, but also the cognitive science world model, with its interpretation of similarity, but not the identity of the language definitions of the phenomena described, in this case manifestations of diseases, having a fuzzy character.

The fuzzy of the world is particularly pronounced fuzziness of the changes occurring in the body of the sick person. At the same time, in the practice of assessing clinical manifestations, there is a fuzzy of the concepts used (attributes) themselves, and of referring them to a certain class. In the process of chronic disease symptoms, as a rule, undergo change, the vagueness of which is differently estimated by each observer of the patient's doctors. In addition, personal characteristics of the patient influenced the manifestation of disease and its dynamics. Thus, fuzzy is determined by the clinical patterns, the variety of known descriptions, the age characteristics of the patient, the knowledge of the expert physicians (involved in the creation of intellectual systems), and their subjective preferences based on past events that have been observed in the past and known from other studies.

The initial fuzzy of descriptions of the observed signs are determined by a number of reasons: transitional states of pathological changes, differences in the manifestations of the disease at different ages of patients within a particular age interval (for example, 1 to 3 years, 4 to 6 years), a rare opportunity to observe patients with orphan diseases. This leads to varying degrees of uncertainty in assessments that characterize the severity of clinical manifestations.

According to Zadeh's granulation principle [8], the degree of granulation of information must correspond to the permissible level of inaccuracy in solving a particular problem, the ability to operate data, knowledge at various levels of detail. The term "granulation" encompasses the processes of composition (formation of larger granules) and decomposition (formation of smaller granules). In this paper, a hierarchy of attributes is used that corresponds to the concept of granularity within the subjective-objective evaluation of characteristics by physicians.

The additional complexity of the process of formation of a textological card is connected with the combination of medical knowledge from publications in different languages and descriptions relating to different ethnic groups with their peculiarities of external manifestations. However, in general, textological cards provide a high-bulk view of the manifestations of the disease in different age periods of life. And the cognitive scientist must be able to choose or reject individual sources.

The structuring of the fragments of the text includes the compilation of a dictionary of signs of different levels from the lower to the meta signs (for example, the contractures of the joints of the hand the multiple contractures of the joints of the limbs multiple dysostosis), taking into account their expression according to the fuzzy linguistic scale. The severity of signs is also represented by a linguistic scale, which includes 7 levels: very strong, strong, moderately, weakly, very weakly, normal value (absence of a pathological sign), and also -1 - impossibility of the manifestation of a pathological sign at a given age. Such a variant of the scale makes it possible to distinguish normal signs characterized by 0, and absence of sign (-1).

At the second stage of knowledge extraction, the cognitive scientist discusses textological cards with the expert. Then, the final variant of the description of the clinical manifestations of the disease, including the level of expression of signs and experts' certainty factors, is chosen. A hybrid structure of process of knowledge extraction by a cognitive scientist from texts and from an expert is presented in Figure 1. As a basis, we took the scheme of extracting knowledge from special texts in the book of T.A. Gavrilova and V.F. Khoroshevsky [7]. Let us consider the existing differences. In the proposed variant, the cognitive scientist structures the text using a linguistic scale. Than the expert, analyze descriptions of signs by different authors in a textological card and, using his knowledge, forms (synthesizes) the final version of the set of symptoms for a specific disease, which are accompanied by certainty factors in the diagnostic significance of these signs, taking into account their severity.

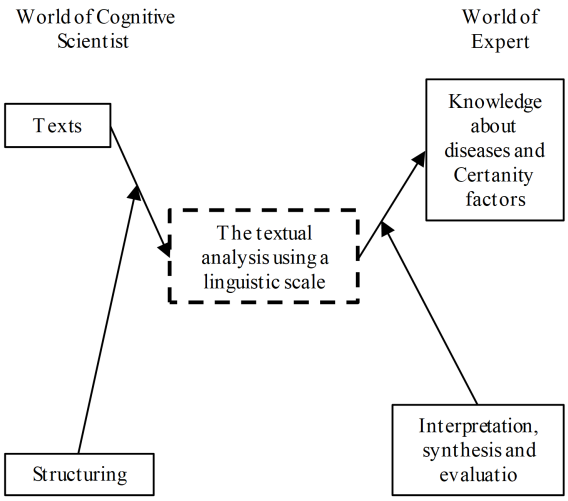

Fig. 1. Scheme of complex knowledge extraction by a cognitive scientist - from texts and from an expert

IV. Work with expert

The views of physicians of different specialties complement each other. Their experience and knowledge are reflected in publications. It allows to represent a multidimensional picture of the disease in a textological card. This is especially important for orphan metabolic diseases due to the fact that there is a gradual, but constantly increasing damage to various morpho-functional systems of the body. And different physicians can have their own viewpoint on the intensity of this process. Also, this is due to visual observation or

the use of special instruments. And for the evaluation of mental disorders, various psychological tests can be used. The cooperation of experts in conjunction with the questions of the cognitive scientist, including probing, allows in the process of discussion to gain additional knowledge reflecting, perhaps, the intuitive representations of medical experts.

In the process of analyzing textological cards, the cognitive scientist specifies the value of individual features in the expert. In addition, he uses various types of questions to determine the degree of reliability of the cognitive with respect to the meaning of the symptoms and their severity, especially when excluding certain symptoms encountered in literary sources. Participation of two experts allows to reveal the lack of information that can be realized in the communication process "cognitologist - two experts".

Working with two experts is particularly important in genetic diseases. As mentioned above, these are rare diseases. Therefore, each expert has in-depth knowledge of different diseases in patients of various ages. A joint dialogue with the cognitologist of two experts can be called in this case the "complementarity principle", as suggested by Niels Bohr in quantum physics. An analogue in the differential diagnosis of hereditary diseases was the group extraction of knowledge from medical experts when creating an expert system [9]. However, the new system is distinguished by the inclusion of the concept of symptom severity and descriptions by age groups, which allows to take into account the progression of the disease course.

Also, when extracting knowledge for the system being created, in the course of a collaboration discussion between the cognitive scientist and the experts, the issue of attribute connections was decided. The importance of the second expert's participation consisted in revealing in the course of the discussion of implicit relationships of attributes, primarily associative ones. In the older age groups, new connections appear, due to the appearance of previously absent features. Regardless of the participation of the cognitive scientist, in this case, the situation was simulated with precedents, which were actively exchanged by experts.

Certainty factors play an important role in the confirmation phase of diagnostic hypotheses. From the standpoint of cognitive linguistics, they reflect both deliberate and intuitive notions of the medical expert. For actuality this process, the expert was shown textological cards. This approach is based on the idea of psychosemantic methods of reconstructing the implicit (deep) knowledge inherent in the subject, which he may not realize, but which are actualized in the "mode of use" [10].In the diagnostic thinking process, an intuitive-shaped component is of great importance for an experienced physician. Therefore, in a dialogue "cognitive scientist - physician-expert" measure of confidence in each of the signs in the age range under special discussion.

V. Conclusion

The hybrid system of knowledge extraction based on the synthesis of textological analysis and expert knowledge makes it possible to extract more useful information for differential diagnostics. At the same time, work is accelerating at the stage of extracting expert knowledge in the process of analyzing prepared textological cards, since the expert receives structured information, previously selected from various literary sources.

Thus, the textological cards formed for each differentiated nosological form of the diseases allow us to visually present to the expert knowledge of the clinical manifestations of the disease with their peculiarities. Comparing them with each other and with their own ideas, the expert forms an integrated description of the disease, supplemented by confidence factors that allow him to assess his measure of confidence in certain characteristics.

References

[1] J.M. Saudubray, H. Ogier de Baulny, C. Charpentier. Clicnical approach to inherited metabolic diseases. Inborn metabolic diseases. Diagnosis and treatment. 3th ed., J. Fernandes, J.M. Saudubray, G. Van den Bergue Eds., Berlin, Springer-Verlag, 2000, pp.3-41.

[2] A.S. Starostin, I.M. Smurov, M.E. Stepanova. A production system for information extraction based on complete syntactic-semantic analysis. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue (2014, June 4-8), 2014, vol. 20, no. 13, pp.659-668.

[3] M.Yu. Khachay, N. Konstantinova, A. Panchenko, D. Ignatov, V.G. Labunets. Analysis of Images. Social Networks and Texts: 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9-11, 2015. Springer, 2015, 460 p.

[4] B.A. Kobrinskii. Nechetkost' i faktory uverennosti verbal'nykh i vizual'nykh ekspertnykh znanii [Fuzzy and certainty factor of verbal and visual expert knowledge]. *Nechetkie sistemy, myagkie vychisleniya i intellektual'nye tekhnologii (NSMVIT- 2017): Trudy VII vserossiiskoi nauchno-prakticheskoi konferentsii. V.2t.*[*Fuzzy systems, soft computing and intelligent technologies (NSMVIT-2017): Proceedings of the VII All-Russian Scientific and Practical Conference. In 2 vol.*], 2017, vol. 1, pp. 83-91

[5] B.M. Velichkovsky, M.S. Kapitsa. *Psikhologicheskie problemy izucheniya intellekta* [*Psychological problems of studying intellect*] in *Intellektual'nye protsessy i ikh modelirovanie* [*Intellectual processes and their modeling*], E.P. Velikhov and A.V. Chernavskii Eds., Moscow, Nauka, 1987, pp. 120-141.

[6] S.A. Sirotko-Sibirskiy. *Smyslovoe soderzhanie teksta i ego otrazhenie v klyuchevykh slovakh* [*The semantic content of the text and its reflection in key words*], Abstract of the dissertations ... candidate of philological sciences, Leningrad, 1968.

[7] T.A. Gavrilova, V.F. Khoroshevsky. *Bazy znanii intellektual'nykh sistem* [*Bases of knowledge of intellectual systems*], St. Petersburg: Peter, 2000, 384 p.

[8] L.A. Zadeh. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy sets and systems, 1997, vol.90, pp.11-127.

[9] B.A. Kobrinskii. Retrospektivnyi analiz meditsinskikh ekspertnykh sistem [Retrospective analysis of medical expert systems]. *Novosti iskusstvennogo intellekta* [*News of Artificial Intelligence*], 2005, no 2, pp. 6-17.

[10] V.F. Petrenko. Lichnost' cheloveka osnova ego kartiny mira [Personality of man - the basis of his picture of the world]. *Rossiiskaya assotsiatsiya iskusstvennogo intellekta* [*Russian Association of Artificial Intelligence*], 1997, pp. 9-24.