Recognition of Folklore Texts and Author's Poems Using Classification Trees and Neural Networks

Ludmila Shchegoleva, Aleksandr Lebedev, Nikolai Moskin
Petrozavodsk State University
Petrozavodsk, Russia
schegoleva@petrsu.ru, perevodchik88@yandex.ru, moskin@petrsu.ru

***Abstract*—The paper deals with the problem of recognition of folklore and literary ("pseudo-folklore") texts. The problem is considered on the one hand by the example of poems of N. A. Klyuev and A. K. Tolstoy and on the other hand of folklore songs of Northern Russia (Karelia) written in late 19th – early 20th century and Luga songs. It is suggested to describe the texts of poems and songs by means of graph-theoretic models which bring to light a set of syntagmatic and paradigmatic relations between the words of texts. It is suggested to compare the graphs of songs and poems by means of a number of numerical characteristics, which were analyzed using classification trees and neural networks. The calculations are performed by means of information system "Folklore" and R program.**

I. Introduction

The variety of connections between folklore and literary tradition can hardly be doubted. Many of the prose writers and poets drew their inspiration in folklore texts, bringing to their works both folk wisdom in general and the system of images inherent in folk art. At the same time, the issue of distinguishing folklore and literary works remains open, if we are talking directly about the analysis of the artistic text and the events presented in it. In general, problem of folklore stylization was considered in the works of M. M. Bakhtin [1] and E. G. Muschenko [2], and also in later studies, where the texts of certain authors and their attempts of folklore stylization were analysed. However, the research presented in these works deals primarily with the issues of stylization under the folklore text at the lexical level of the language, while the syntactic and semantic components of the texts are practically not considered. It becomes necessary to identify the criteria that make it possible to distinguish the folklore texts themselves from those which are only stylized as folklore, as well as the possible automation of the separation process of these two text groups. During the analysis, an attempt was made to analyze the semantic structure of texts based on graph-theoretic models, including syntagmatic and paradigmatic relations between objects, and demonstrating the relationship between these objects and their actions.

The object of the current study were four groups of texts. Two of them are folklore texts, "The Songs of Zaonezhie" [3] and "Luga Songs" [4]. As an additional comparative material, the texts of N.A. Klyuev [5] and A. K. Tolstoy [6] were used. In their poetic texts "folk" elements and closeness to Russian culture have traditionally attracted the attention of both readers and researchers of the poetic text. The closeness to the folklore tradition was manifested both in the semantic and in the poetic aspect, which brings their poems closer to the samples of folk art.

II. Graph-theoretic models of the texts and their numerical characteristics

The graph-theoretic model of the text is a 4-tuple $G=(V, E, \alpha, \beta)$, where:

- V is a finite set of vertices;
- $E \subseteq V \times V$ is the set of edges;
- α: $V \to L_V$ is a function assigning labels to the vertices;
- β: $E \to L_E$ is a function assigning numbers to the edges.

Similar model is described on the material of the folklore songs of Zaonezhie XIX – early XX century [7]. The vertices of the graph correspond to objects such as characters, natural phenomena, animals, household items, etc. Between objects we set connections, which are divided into two groups: local and global (these are edges of the graph). Local connection exists in the case when the relation between objects are expressed directly in the text (a verb, a verb form or an adjective). Therefore each local connection can be match its serial number of appearance in the text. Global link is established if the relationship between objects is not expressed in the text as word forms, but there are recognized by reader. If you combine the same objects in one vertex, then a similar structure can be represented in a single graph of song.

Here L_V is the set of labels and attributes of objects defined in subject area. L_E is the set of nonnegative integers from 0 to n, where n is the total number of local links in the text. If the connection e belongs to a global type then $\beta(e) = 0$. Else $\beta(e)$ coincides with its serial number of appearance in the text. Note that if e_i and e_j are local links, then $\beta(e_i) \neq \beta(e_j)$.

With the help of the information system "Folklore" [7] the following numerical characteristics were calculated from the source texts and their corresponding graphs:

- *word* – number of words in the text;
- *string* – number of lines in the text;
- *vertex* – number of vertices in the graph;
- *edge* – number of edges in the graph;
- *max* – the maximum degree of the vertex in the graph;

- *link* – connectivity parameter of the graph. It is found as the ratio of the number of pairs of connected vertices to the number of edges in the corresponding full graph with the same number of vertices;
- *a* and *b* – the coefficients of the hyperbolic regression. The objects in the studied texts are not equivalent between themselves. Two main objects often dominate in the story (vertices with the maximum degree). The remaining objects are secondary, they appear in the text less (most one or two times). If we juxtapose each node of the graph and its degree and sort them in descending order, we get graphics that you can interpolate with the hyperbolic curve $y=a/x+b$;
- *global* – percentage of global links (which reflect the paradigmatic links in the text) of the total number of links.

III. The construction of classification trees with R program and their interpretation

In this paper we continue analysis of approaches to discrimination of folklore and literary texts [8], [9]. We enlarge the collection and investigate two classification methods: decision trees and neural networks.

Classification problem is the assignment of some sort of output value to a given input value. In our case we have set of numerical characteristics were calculated from the source texts and their corresponding graphs as an input and information about the attachment of each text to one of four groups as an output. Table I contains a fragment of the input values for graph-theoretic models of five Klyuev's poems. The output vectors were formed in accordance with methods used.

At first we used the binary decision trees as a model for classification [10]. The sample population consisted of 210 observations and nine explanatory variables. The explanatory variables *word*, *string*, *vertex*, *edge*, *max* are discrete, variables *link*, *a*, *b*, *global* are continuous. The predicted variable takes four values: 1 for Klyuev's poem, 2 – folklore songs of Zaonezhie, 3 – A. K. Tolstoy's poetic works, 4 – Luga songs.

TABLE I. Input values for classification task (parameters of graph-theoretic models of five Klyuev's poems)

word	string	vertex	edge	max	link	a	b	global
106	27	21	20	11	0,095	26,067	0,237	0,450
59	16	16	15	7	0,125	22,957	1,399	0,667
77	17	12	12	7	0,182	24,534	1,989	0,833
55	16	12	13	8	0,182	29,406	0,729	0,462
71	16	11	10	7	0,182	34,495	-0,379	0,700
…	…	…	…	…	…	…	…	…

All decision trees were built by means of routings of the R package "rpart". We built eight decision trees. One decision tree was built to classify four groups of texts. The precision of classification equals 70%. The decision tree identifies some features characterized every group and distinguish it out from others.

The next four decision trees were built for each group separately to get the group's differences from the rest of the texts in the aggregate. Two more decision trees were built to separate Klyuev's poem from A. K. Tolstoy's poetic works in the group of literary texts and to separate folklore songs of Zaonezhie from Luga songs in the group of folklore texts. And finally one decision tree was built to distinguish folklore texts and literary text. The precision of classification equals 90%. In this case the predicted variable takes two values: 0 for literary texts (Klyuev's poem and A. K. Tolstoy's poetic works), 1 for folklore texts (folklore songs of Zaonezhie and Luga songs).

In accordance with decision trees, it is clear that the key parameter, paramount in terms of fragmenting texts into groups, is the *link* parameter. The low connectivity is the parameter that distinguishes Klyuev's verse texts (43 out of 50 texts, 86%) from the rest of the groups. Moreover, a number of Klyuev's poems (18 out of 50, 36%), in addition to low connectivity, are characterized by a small parameter *a* ($a<12,78$), which uniquely distinguishes these texts from all others. Weak connection between objects can be considered as distinctive feature of the analyzed works – Klyuev's poetry is filled with lyrical images and characters, mentioned only once in the text of the work and therefore have a syntactic connection with only one other object, or do not have such connections in an explicit form at all.

The texts of A. Tolstoy were divided approximately in half by the *link* parameter, which made it difficult to determine their specifics.

For folklore songs of Zaonezhie, it is necessary to highlight key parameters such as high connectivity (the density of links is typical for folklore texts – often the relationship is built between one or two lyrical subjects, which, moreover, are also mentioned in text repeatedly thanks to the use of synonyms and personal pronouns, as well as a small percentage of global connections – for folk texts, actions are realized in verbs of the text (the local connection), in comparison with literary works (in this groups of texts actions are often implied by the author or exist only in the mind of the reader).

The set of Luga songs characteristics is similar to the set of folklore songs of Zaonezhie. It is interesting that on the basis of the data obtained, it was possible to identify a significant set of parameters that distinguish groups 1 and 3 (literary texts) from groups 2 and 4 (folklore texts): $link\geq 0,129$, $global<0,4451$, $a\geq 16,8$. Under these parameters, 92 out of 110 folklore texts (84%) fall and only 15 out of 100 literary texts (15%).

It is significant that, in general, *word* and *string* parameters proved to be of little importance, since works of different volume were involved in analysis in all four groups of texts. Curiously, certain sets of parameters included examples from three or more groups of texts. In particular, a set of parameters ($link<0,129$, $a>12,78$, $30,5>string\geq 15$, $max<13,5$, $edge\geq 16,5$) correspond to at least one example from all four groups of texts. It is interesting to observe not only typical texts for a particular group, but, on the contrary, pay close attention to

those works that significantly differ from the whole group and are mixed in terms of parameters with the texts of other groups. This may indicate a very high-quality stylization for folklore (in the case of literary texts) or the of a certain folklore text for works written or performed on the territory in general.

As a result we formulated several classification rules.

IV. Application of Neural Networks

The second method of classification we used was neural network [11]. The neural networks were built by means of routings of the R package "nnet". The output vectors were designed as follows. The length of the vectors equals amount of input texts. The first vector corresponds text of Klyuev's poems. For every text from input sample if the text is a Klyuev's poem, then the corresponding element of the first vector has the value 1, otherwise – 0. Similarly, the remaining vectors were constructed. The second vector corresponds folklore songs of Zaonezhie, the third vector – A. K. Tolstoy's poetic works, the fourth – Luga songs.

The sample population consisted of 198 observations. We excluded three observations from each group of texts to make the test collection. We constructed single-hidden-layer neural network with 9 input units and 4 output units. The best results were obtained for 20 units in the hidden layer. The precision of classification on the training collection equals 100%. At the same time the precision on the test collection is less than 50%.

The second experiment was carried out for the classification of folklore texts and literary texts. We constructed single-hidden-layer neural network with 9 input units and 1 output unit. The response 0 on the output unit corresponds literary text, 1 – folklore text. For the training collection the precision of 100% was obtained for 11 units in the hidden layer of neural network. The precision on the test collection was in 50–80%. The increase in the number of units above 11 in the hidden layer also gave 100% precision for the training collection, but did not lead to an increase in precision on the test collection.

V. Conclusion

This paper continues the study of distinction of folklore and literary ("pseudo-folklore") texts. Two approaches to classification based on decision trees and neural networks were investigated.

References

[1] M.M. Bakhtin, *Questions of Literature and Aesthetics*. Moscow, 1975.

[2] E. G. Muschenko, V. P. Skobelev, L. U. Kreichik, *Poetics of the Tale*. Voronezh, 1978.

[3] R.B. Kalashnikova, *Besedy and besednye songs of Zaonezhie in the second half of the 19th century*. Petrozavodsk, 1999.

[4] *Songs of the Gorodensky Choir* / Compilation, preface, notation of tunes by E. E Vasilieva. Novgorod, 1990.

[5] N.A. Klyuev, *Heart of the unicorn*. Verses and poems. Saint-Petersburg, 1999.

[6] A.K. Tolstoy, *Collected works*: in 4 volumes. Moscow, 1969. Vol. 1.

[7] N.D. Moskin, *Theoretical-graph models of folklore texts and methods of their analysis*. Petrozavodsk, 2013.

[8] N.D. Moskin, A.G. Varfolomeyev, A.A. Lebedev *Comparative analysis of textual structure of author's poems and folklore songs by graph-theoretic models* // Papers from the Annual International Conference «Dialogue-2017. Computational Linguistics and Intellectual Technologies». Moscow, May, 2017. Issue 16. Vol. 1. 15 p. Web: http://www.dialog-21.ru/media/3977/moskinndetal.pdf.

[9] N.D. Moskin, A.A. Lebedev *Classification of poetic and folklore texts by the method of discriminant analysis* // Proc. of the Conference «Humanitarian Education and Science in a Technical University». Izhevsk, Oct. 2017. P. 463-467. Web: https://sites.google.com/view/academicconference/

[10] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone *Classification and Regression Trees*. Wadsworth, Belmont, Ca, 1983.

[11] B.D. Ripley *Pattern Recognition and Neural Networks*. Cambridge, 1996.