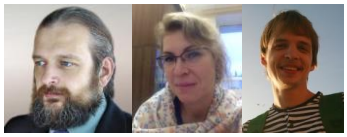
The approaches to the construction of a joint ("two-dimensional") classifier of genre types and stylistic colouring of poetic texts

V Barakhnin[1, 2*], O Kozhemyakina[1], I Pastushkov[1]

[1]Institute of Computational Technologies of SB RAS, Lavrentiev av., 6, 630090, Novosibirsk, Russia
[2]Novosibirsk State University, Pirogov str.,1, 630090, Novosibirsk, Russia

**Corresponding author's e-mail: bar@ict.nsc.ru*



Abstract

In this paper we propose the algorithm of automated definition of the genre type and semantic characteristics of poetic texts in Russian. We formulated the approaches to the construction of a joint ("two-dimensional") classifier of genre types and stylistic colouring of poetic texts, based on the definition of interdependence of the type of genre and stylistic colouring of the text. On the basis of these approaches the principles of formation of the training samples for the algorithms for the definition of styles and genre types were analysed. The computational experiments with a corpus of texts of the Lyceum lyrics of A.S.Pushkin were implemented, which showed good results in determining the stylistic colouring of poetic texts and sufficient results in determining the genres. The proposed algorithms can be used for automation of the complex analysis of Russian poetic texts, significantly facilitating the work of the expert in determining their styles and genres by providing appropriate recommendations.

Keywords: automated analysis, computational experiments, "two-dimensional" classifier, genre, style

1 Introduction

In the tasks of automated text analysis in natural language, the problem of determination of their genre and stylistic characteristics is determined. The researcher can get this problem in a wide range of situations: from the problems of automation of the complex analysis of poetic texts, for which the type of genre and stylistic characteristics are the important attributes used in determining of the impact of lower levels on higher levels of the verse (see for example [1]), to the tracking of messages in social networks to identify the terrorist threats, the determination of marketing preferences of buyers, etc.

The researches in the field of automated determination of the genre type of texts were started recently – in early 2010-ies. So, in work [2] the algorithms of determination of genre types of odes, songs, epistles, elegies and epitaphs are based on the works of English poets-sentimentalists of the XVIII century. The time period in this study was not chosen by chance: in the poetry of the XVIII century the classicism with its strict genre rules dominated, and this greatly facilitated the development of algorithms.

The paper [3] describes the method of text classification (for certain genres and authors) based on the analysis of statistical regularities of letter distributions, i.e. the probabilities of occurrence of letters and letter combinations, along with this a solution is found without the "invasion in the sphere of literature, i.e., without the analysis of syntax, literary techniques and patterns of character interactions". However, in [4], the authors build an original counterexample to the statistical method of identification that shows the necessity of using, at least, the methods of morphological analysis.

As for the automation of determination of stylistic characteristics of the texts, we don't know the researches in this area, at least for the texts in Russian. Thus, our researches on computer joint definition of the type of genre and stylistic colouring of Russian texts are of a pioneer nature.

In the present work we implement to develop the approaches to the construction of a joint classifier of the types of genre and stylistic colouring of poetic texts as well as we made the comparative analysis of algorithms of determination of these characteristics. Our purpose is not the creation of new theories of genre and stylistic relationships within literary works but the development of the analyser that allows to correlate correctly the stylistic colouring of the text with its genre identity what has relevance for researches in the field of Informatics, because we are talking about the tools used not in the strictly linguistic space.

2 The choice of training samples

While we built the joint ("two-dimensional") classifier of genre types and stylistic colouring of texts, we took into account that the classifier itself is a multidimensional structure, based on the totality of parameters, which define the object of study. When we construct the multidimensional classifiers associated with such difficult (for unequivocal definition) categories like genre and style, the phased development of each analysis parameter is required in order to exclude possible errors and the variability of results. Such classifier is created for the first

time (at least for texts in Russian). This task requires the synthesis of a vast empirical material, so we decided to stay on the classification of lyrics of A.S.Pushkin of Lyceum period, as it has the most strict genre forms, stylistic unity, and adherence to the rules of grammar of that period. In turn, the usage of the Lyceum lyrics as material for the creation of training samples is justified by stylistic dimension, since the stylistic differentiation of lexemes is the development stage of the classifier.

Genre types formed the basis of the classifier: along one axis we have placed the genre types in order of ascending "the sublimity" and along another axis - the traditional styles (see Table 1).

TABLE 1 The statistics on the genre and stylistic compliance

	High	Neutral	Low
Ode	4	-	-
Parable	1	1	-
Madrigal	4	-	-
Epistle	-	55	5
Idyll	-	2	-
Elegy	-	37	-
Romance	-	1	-
Ballad	-	3	-
Epigram	-	-	18
Anecdote	-	-	1

In general, the style of a text is determined by the most "low" of its lexemes and is characterized by the lexemes much more than a genre, although, in our experiment, in particular, and to the global literary categories processes and, in general, the number of genres is much more than a number of styles. This complicates the choice, as because of direct factors, and also because of given training sample of works.

3 Description of the numerical experiment

For the experiment we used the above-described massive of the texts of Pushkin's lyrics of Lyceum period, comprising 121 poems, marked by an expert on genres and styles.

We used the standard method of support vectors machine (support vector machine) [5] with a linear core and the RBF nonlinear core, in addition, for comparison, the results of calculations were carried out with the use of neural network based on multilayer perceptron [6]. When training the dictionary of all used words was created, except service words, and each text was coded by sequence of the symbols 0 and 1 corresponding to the dictionary in word order: 0 was set if a word is not in the text, 1 – if the word is in the text. Also we used the linear regression to determine the styles, our hypothesis was that as styles can be unambiguously ranked: low – 1, neutral – 2, high – 3, the regression can give a numeric value which will be close to the value of the style, and a divergence with the value will be a mistake. The experimental results are following (see Table 2): we calculated the average, the minimum, and the maximum of the proportion of correct predictions of the method with 100 runs, the sample is divided into 80% training and 20% test, the division into which is random every time, each run is independent from the previous ones (algorithm was implemented in the language python using the library scikit-learn). As it is difficult to rank a neutral along with the others, than in each method there is the experiment with it and without it.

As can be seen from the obtained data, the high style is not practically recognized – probably because of its insufficient representation in the sample. The method of support vector machines is the best in this case. It is worth to note that in the case of non-linear core the high style was recognized, but by the common parameters, the case of linear core is better than the multilayer perceptron and logistic regression.

TABLE 2 Experiment with the definition of the style

	Average value	Max	Min
SVM, neutral is ignored	0.76	0.92	0.58
SVM	0.80	0.96	0.57
SVM, RBF core	0.62	0.85	0.11
Multilayer neural network	0.77	0.96	0.46
Logistic regression	0.76	0.96	0.46
Linear regression, neutral is ignored	0.70	0.82	0.45
Linear regression	0.70	0.45	0.58

	High	Neutral	Low
SVM, neutral is ignored	0.0	0.86	0.72
SVM	0.0	0.86	0.70
SVM, RBF core	0.10	0.75	0.13
Multilayer neural network	0.0	0.96	0.33
Logistic regression	0.0	0.85	0.72
Linear regression, neutral is ignored	-	-	-
Linear regression	-	-	-

Similarly, we carried out the experiment on definition of the genre (one series of experiments was carried out under the simplified scheme, when the historical elegy and philosophical ode was not seen as separate genres). From Table 3 it is seen that the definition of genre has fared worse than the definition of styles as each genre is represented by a relatively small number of samples. The lexical signs are not enough for genres, we need poetic features (rhyme, size, number of accented syllables) which should strengthened, for example, with the AdaBoost algorithm [6].

TABLE 3 Experiment with the definition of the genre

	Average value	Max	Min
SVM, simplified types	0.45	0.65	0.22
SVM	0.45	0.65	0.27

4 Conclusions

The paper proposes the approaches to the construction of a joint ("two-dimensional") classifier of genre types and stylistic colouring of poetic texts, based on the definition of interdependence of the type of genre and stylistic colouring of the text. On the basis of these approaches we analyse the principles of formation of the training samples for the algorithms to define styles and genre types. We implement the computational experiments with a corpus of texts of the Lyceum lyrics of A.S.Pushkin, which showed good results in determining the stylistic colouring of poetic texts and sufficient results in determining the genres. Thus, the proposed algorithms showed their efficiency and can be used for automation of the complex analysis of Russian poetic texts, significantly facilitating the work of the expert in determining their styles and genres by providing appropriate recommendations.

Acknowledgments

Work is executed with partial support of the Presidium of RAS (project 2016-PRAS-0015) and of the Presidential programme "Leading scientific schools of RF" (grant 7214.2016.9).

References

[1] Barakhnin V, Kozhemyakina O 2012 About the automation of the complex analysis of Russian poetic text *CEUR Workshop Proceedings* **934** 167-71 (*In Russian*)
[2] Lestsova M 2014 The determination of the core and the periphery of the genres of odes, songs, epistles, elegies and epitaphs on the works of English poets-sentimentalists of the XIX century *Bulletin of the Chelyabinsk State Pedagogical University* **4** 196-205 (*In Russian*)
[3] Orlov Yu, Osminin K 2010 The definition of the genre and the author of a literary work by statistical methods *Applied Informatics* **26**(2) 95-198 (*In Russian*)
[4] Orlov Yu, Osminin K 2012 *Methods of statistical analysis of literary texts* Editorial URSS: Moscow (*In Russian*)
[5] Cristianini N, Shawe-Taylor J 2000 *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* Cambridge University Press
[6] Freund Y, Schapire R E 1999 A Short Introduction to Boosting *Journal of Japanese Society for Artificial Intelligence* **14**(5) 771-80