УДК 004.89

Boris Melnikov[1], Svetlana Pivneva[2]

[1] Samara National Research University, Samara, Russia
[2] Togliatti State University, Togliatti, Russia

ON THE MULTIPLE-ASPECT APPROACH TO THE POSSIBLE TECHNIQUE FOR DETERMINATION OF THE AUTHOR'S LITERARY STYLE*

ABSTRACT

We consider in this paper our approach to the determination of the author's literary style. The preliminary results for both Russian and English literary texts are very promising. In fact, only some statistical methods for the words were considered in the previous known papers. Our approach includes the following stages: "manual" obtaining some characteristics (descriptions) of the considered texts; obtaining the "simple" representation of the text; obtaining the representation of the considered text using so called π-subclasses; using special archiver for all the obtained representations; applying multiheuristical approach for solving appeared discrete optimization problems; using special clustering algorithm.

KEYWORDS

Multiheuristical approach, special clustering algorithm, literary style.

Мельников Б.Ф.[1], Пивнева С.В.[2]

[1] Самарский национальный исследовательский университет имени академика С.П. Королева, г. Самара, Россия
[2] Тольяттинский государственный университет, г. Тольятти, Россия

МУЛЬТИПЛИКАТИВНЫЙ ПОДХОД К ТЕХНИКЕ ДЛЯ ОПРЕДЕЛЕНИЯ ЛИТЕРАТУРНОГО СТИЛЯ АВТОРА

АННОТАЦИЯ

В статье рассматривается наш подход к определению литературного стиля автора. Получены удачные предварительные результаты для русского и английского литературных текстов. Существует несколько статистических методов для определения литературного стиля, известных авторам. Наш же подход включает в себя следующие этапы: "ручное" получение некоторых характеристик (описаний) рассматриваемых текстов; получение "простого" представления текста; получение представления текста с использованием так называемых π-подклассы; с помощью специального архиватора для всех полученных представлений; применяя мультиэвристический подход для решения появившихся задач дискретной оптимизации; использование специального алгоритма кластеризации.

КЛЮЧЕВЫЕ СЛОВА

Мультиэвристический подход; специальный алгоритм кластеризации; литературный стиль.

Introduction and Preliminaries

We consider in this paper our approach to the determination of the author's literary style. For now, we are working with Russian and English classical texts and applying our algorithms only to them. In the nearest future, we are going to consider also German classical texts.

For now, we only began such works; we are accumulating the statistical data. Therefore we will not publish the obtained concrete preliminary results. However, these preliminary results are very promising, and we hope to publish them in the next papers. And it is important to remark, that these preliminary results are promising for both Russian and English literary texts. Thus, our paper can be considered as the detailed description of our approach to this problem (which can be considered as a problem of artificial intelligence),

* **Proceedings of the XI International scientific-practical conference «Modern information technologies and IT-education» (SITITO'2016), Moscow, Russia, November 25 - 26, 2016**

and such approach was already applied in some other problems of artificial intelligence.

Despite such questions (i.e., the automatic determination of the author's literary style) were considered since at least 1916 ([1]), author does not know the papers where the literary style is determined by the investigation of grammatical structures.

However, there were a lot of papers where a lot of approaches to considering grammatical structures were described; but after [2] (1993), there was very likely no good reviews of such papers. However, author does not know papers where the determination of the author's literary style was considered using grammatical structures.

In Russian*, such possible techniques were reviewed in [3,4]. However, only some statistical methods were considered in the techniques reviewed there. Let us remark that in the well-known work [5] (also related to the Russian literary texts), also only some statistical methods for the words were considered. Below, we shall call such methods by "usual" ones.

And we propose in this paper a multi-aspect approach to this problem. Our approach includes the following stages:

A “manual” obtaining some characteristics (descriptions) of the considered texts. Remark that some other aspects of our approach can (i.e., its stages considered below) be considered as automatically obtaining such characteristics. Applying the "manual" characteristics to the special representation of considered texts.

Obtaining the “simple” representation of the text (i.e., representation for “usual” algorithms).

Obtaining the representation of the considered text using so called π-subclasses. It is a special description of the grammatical structures. This description can be considered as an alternative to the description using Chomsky hierarchy. We already applied such description in some problems of formal languages theory ([6,7]), and began to apply it in this problem. Most of author's results connected with such representation were published in Russian; in English, some its applications (for the representation of some formal languages, at first, simple programming languages) and also the references can be found in [7,8].

Using special archiver for all the obtained representations. For now for its description (i.e., for algorithms of archiving and some its practical results), we have a master thesis only [9]; we are going to publish these results in the nearest future. Let us remark, that the goal of considering algorithm of this archiver is not the data compression, but automatically obtaining some characteristics of considered texts.

For each used representations, applying multiheuristical approach (solving appeared discrete optimization problems) to special comparing of two long string representing two considered texts. See [10] for the detailed description of such approach.†

Using special clustering algorithm, i.e., our version of hierarchical clustering algorithm. Clustering can be used to obtain some texts which seem to belong to the same author. This version was of hierarchy clustering algorithm was published only in Russian [11].

In the next sections, we shall describe some stages of our approach (i.e., its components which are interested for the subject of this paper) some more detailed.

A “manual” obtaining some characteristics of the considered texts

This section is “most informal”, and, therefore, we cannot describe it detailed. Really, the possible characteristics (i.e., descriptions) of the considered texts strongly depend on the used language. Therefore we can consider the obtaining such descriptions as a new sub-problems of artificial intelligence, which are similar to the problems of obtaining knowledge from the expert and formalization these knowledge in some expert systems (we mean rule-based approach here).

Thus, our characteristics could be very different for different considered languages. We shall consider briefly only two used examples of such characteristics used in our computer programs for Russian literary texts.

One of them is using phraseological stock phrases. However by [3],‡ “much more often, in the author’s texts there are no pronounced words and word constructions, no prominent stock phrases”. But we already began to analyze this characteristic.

And the second possible characteristic is the relative frequency of using loan words.

In both such cases, we need some dictionaries (data bases of corresponding text structures) for obtaining such characteristics. Moreover, in the first case, we need something like data base using our representation of the grammar structure.§ But let us remark, that working with such data bases cannot be

* We mean here both Russian scientific papers and Russian language for determination of the author's literary style.

† After [10] (2005), author has published some other papers on this multiheuristical approach. But it is not important for the subject considered here, i.e., for the possible techniques for determination of the author's literary style

‡ Remember that [3] is devoted to Russian texts.

§ See also Section 3.

considered as working with "simplest statistical method" which was criticized before.

Obtaining the representation using π-subclasses

As we said before, we use a special description of the grammatical structures, which can be considered as an alternative to the description using Chomsky hierarchy. We have already applied such description in some problems of formal languages theory ([6,7]), and began to apply it in this problem. Most of author's results connected with such representation were published in Russian; in English, some its applications (for the representation of some formal languages, at first, simple programming languages) and also the references can be found in [7,8].

The next example* can be considered as a connection (a "bridge") between using π-subclasses in formal and natural languages. Let us consider grammatical structure of programming language Pascal, which defines compound statement. We shall use special brackets '(and)' to denote iterations and ε for the empty word. The other designations (i.e., A, B, L, μ and Ψ) were defined in [7]. Thus, let

[compound statement] ::= begin [sequence of statements] end

[sequence of statements] ::= [statement] '(;[statement])'

We can define here

L = { begin [sequence of statements] end }

A = { begin '([statement];)' ε }

B = { ε '(;[statement])' end }

Remark that we succeeded in describing the language L without the productions for the non-terminal

[sequence of statements]. And then the language defined by the construction [compound statement] is Ψ*(μ,L). Remark that all the needed conditions (i.e., conditions for Proposition 2 of [7]) hold.

Thus, this example demonstrates a successful applying of π-subclasses. And the same constructions can be used in natural languages (at first, in English). Without strict mathematical theory, the profit of applying π-subclasses can be defined for English infinitive phrases (adjective, adverb or noun ones). Recognizing them, we use the same π-subclasses as we considered in the example for formal programming language.†

And the similar acceptable representation can simply be also applied to the following English phrases: present ones; past participial ones; gerund ones; prepositional ones; absolute ones; appositive ones; and also adjective, adverb and noun subordinate clauses. All this structures can be successfully represented using π-subclasses.‡

A short description of special algorithm of archiving text

In this section, we shall very briefly describe our simple algorithm of the text archiving. As we said before, the goal of considering algorithm is not the data compression,§ but automatically obtaining some characteristics of considered texts.

Thus, let us consider the usual 8-bit letter representation as the 0th stage of the archiving.**

On the 1st stage we pass to the 9-bit representation. Using this passing, we use the high-order bit to designate what we have in other 8 bits: the letter or the number of pairs. Such possible 256 pairs represent the "best pairs" of the input data, i.e., the mostly used pairs in the considered text.††

After that we pass to the 10-bit representation, etc. However, as we said before, the goal is not the best comparing, but automatically obtaining some characteristics of considered texts. It is important to remark, that like dealing with the algorithms of neural networks, we often "do not understand what our algorithm makes", i.e., we obtain the needed characteristics independently on the texts of our computer programs. Thus, we formulated the main difference between "manual" and automatically obtaining the

---

* It can be considered as an modified and improved Example 2 of [7].

† Moreover, we use some simplified construction, because, for example, we do not need here the unlimited enclosing of considered constructions.

‡ To finish this section, let us remark the following thing. Algorithms for working with π-subclasses can include auxiliary algorithms for working with finite automata of special form. (Because π-subclasses can be considered as special representations of regular formal languages.) And the most acceptable short representation of such automata can be constructed by the theory [12]. However, these things are far from our subjects.

§ Or, rather, the goal of this paper is not the data compression. We have obtained some successful results for some groups of archived data. We are going to publish such results in the nearest future.

** Certainly, for most languages we may use 5 bits (i.e., 32 letters) for this 0th stage. This possible alternative (i.e., whether we use the recoding into the 5-bit letter representation and start from it) can be considered as an additional heuristic. Also we can use arithmetical (Huffman) coding (see [13] etc), which also can be considered here as a heuristic. Remark that usually this coding is used after the main algorithm, but we use this coding (a subsidiary transformation) before it.

†† Remember that we are working not only with the given texts, but also with some their representations.

characteristics of the given texts.

In the compressing algorithms, this process finishes when the real compressing ends. And in our case, there is better to choose the number of bits before the starting algorithm; but let us remark, that the best choosing number (i.e., for obtaining best possible characteristics) is usually equal to the number when the compressing ends.

A short description of special clustering algorithm

In this section, we consider a short description of our clustering algorithm; the used notation is standard for such area. Thus, let Rij be minimal distance* between elements belonging to the clusters having numbers i and j, and ri is the maximal distance between 2 elements of cluster number i. The standard clustering algorithms and also heuristical algorithms for clustering quality usually use values

$$f(\min_{i,j}(R_{ij})) + g(\max_{i}(r_i)),$$

where f and g are special increasing functions (depending on concrete problem, on concrete algorithm, etc), and i and j have all the possible values.

Unlike standard algorithms, we use the same formula, but values ri have other sentences. Namely, let:

i be the considered cluster;

m1,...,mn be all the its elements;

lmn be the distance between its elements having numbers m and n;

(M1,..., MN) be some finite sequence of the elements m1,...,mn (certainly, all these elements belong to the considered cluster) and including each of them at least 1 time.

Then we set

$$r_i = \min_{(M_1,\ldots,M_N)} \left(\max_{0<K<N} (l_{M_K M_{K+1}}) \right).$$

Remark that the given definition is not an algorithm, but we can simply obtain some algorithms (for obtaining ri,) constructed on its base.

The practical programming† shows the acceptability of this algorithm in various problems. As we said before, we use it to obtain some texts which seem to belong to the same author.

Conclusion

Let us consider a generalization of previous sections.

By the author's opinion, the main weakness of previous approaches for determination of the author's literary style is the over- passion of syntactical analysis of the texts, and at first (as the main characteristic of the considered texts) storing and processing trivial frequency statistic of the used words. And we focus the main attention to the analysis of the morphological formal constructions of the author's sentences, at first, to the special characteristics of used morphemes. We also are going to continue the described before multiheuristical approach to discrete optimization problems; we mean here its applying to some sub-problems considered in this paper.

Thus, we have considered the brief description of our method for the possible technique for determination of the author's literary style. And, as we said before, we are going to publish some result of practical programming in the next papers.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, соглашение по проекту №16-47-630829.

References

1. A.Markov: On an application of statistical method. – Izvestia Imperatorskoy Akademii Nauk (Russian Academy of Sciences Ed.), 1916, Ser.VI, Vol.X, No.4, 239–239 (in Russian).
2. E.Charniak: Statistical language learning. MIT Press, 1993.
3. http://www.rusf.ru/books/analysis/history.htm (in Russian).
4. http://www.computerra.ru/offline/2000/338/3010/ (in Russian).
5. J.Kjetsaa, S.Gustavsson, B.Beckman, S.Gil: The Authorship of The Quiet Don. – Solum Forlag A.S., Oslo and Humaities Press, N.J., 1984.

* Let us also remark, that we have to have a priory defined metrics there. However, we have indirectly said about such possible metrics in Introduction: we said that we apply our multiheuristical approach solving some discrete optimization problems to special comparing of two long string representing two considered texts. And this comparing forms the needed metrics.

† See some references in [11, 14].

6. O.Dubasova, B.Melnikov: On the generalization of the context-free languages class. – Progtammirovanie (Russian Academy of Sciences Ed.), 1995, No.6, 46–58 (in Russian).
7. B.Melnikov, E.Kashlakova: Some grammatical structures of programming languages as simple bracketed languages. – Informatica (Lithuanian Acad.Sci. Ed.), 2000, Vol.11, No.4, 441–454.
8. B.Melnikov: Some equivalence problems for free monoids and for subclasses of the CF-grammar class. – Number Theoretic and Algebraic Methods in Computer Science, World Sci. Publ., 1995, 125–137.
9. O.Sannikova: A special archiver. Master thesis, Togliatti State University, 2008 (in Russian).
10. B.Melnikov. Discrete optimization problems – some new heuristic approaches. – 8th Int. Conf. on High Performance Computing and Grid, IEEE Comp. Soc. Press Ed., 2005, 73–80.
11. B.Melnikov, E.Melnikova: Clustering situations in real-time algorithms for discrete optimization problems. – Sistemi upravleniya i informacionnie technologii (Institute of Control Problems of the Russian Academy of Sciences Ed.), 2007, Vol.28, No.2, 16–19 (in Russian).
12. B.Melnikov, N.Sciarini-Guryanova: Possible edges of a finite automaton defining the given regular language. – The Korean Journal of Computational and Applied Mathematics, 2002, Vol.9, No.2, 475–485.
13. D.Huffman: A method for the construction of minimum-redundancy codes. – Proc. Inst. Radio Engineers, 1952, Vol.40, No.9, 1098–1101.
14. B.Melnikov, S. Pivneva: Heuristic algorithms of decision-making in humanitarian areas. - News of the Samara centre of science of the Russian academy of sciences, 8 - Samara: Publishing house of the Samara centre of science of the Russian Academy of Science, 2009.

Поступила 14.10.2016

Об авторах:

Мельников Борис Феликсович, профессор, заведующий кафедрой прикладной математики и информатики Тольяттинского филиала Самарского национального исследовательского университета имени академика С.П. Королева, bf-melnikov@yandex.ru;

Пивнева Светлана Валентиновна, доцент кафедры высшей математики и математического моделирования Тольяттинского государственного университета, кандидат педагогических наук, tlt.swetlana@rambler.ru.