Disputed Authorship Resolution Using Relative Empirical Entropy For Markov Chain of Letters in a Text

D.V. Khmelev

25 August 2000

AFFILIATION: Heriot-Watt University, Edinburgh, U.K. and Isaac Newton Institute for Mathematical Sciences, Cambridge, U.K.
POSTAL ADDRESS: 20 Clarkson Road, Cambridge, CB3 0EH, U.K.
FAX NUMBER: (01223) 330508
E-MAIL ADDRESS: D.Khmelev@newton.cam.ac.uk

A new statistical method in the analysis of literary style for disputed authorship resolution is considered here. It was tested in the following experiment.

We take a training set of 304 text samples (novels, stories and short stories) of 82 different authors in Russian. The total size of training text samples for each author exceeds 100,000 symbols. We also take one control text sample per each author. The size of the control sample exceeds 100,000. The total volume of text samples is about 120Mb.

Each control text is considered to be anonymous. Our method determines the true author for 69 control texts. In 3 (resp. 2, 1) cases the true author is second (resp. third, fourth) in the list of pretenders to its own text. Note that the analysis of frequencies of isolated letters guesses the true author for just 2 control texts.

Let us describe the method in detail. Consider a sequence of letter of text as a Markov chain. The matrices of transition frequencies of letters pairs are calculated over all texts for each author. Therefore we know (approximately) the probability of transition from one letter to another for each author. The author of the control text is guessed by the principle of maximal likelihood, i.e., for each matrix we calculate the probability of anonymous text and we choose the author with the maximal corresponding probability. The chosen author is supposed to be the true author.

Suppose we take the logarithm of each probability, change a sign, and divide it by the length of control text; then each of the numbers obtained is called the *relative empirical entropy*. Relative empirical entropy is more convenient for computing than actual probabilities. Besides, the chosen author (who often happens to be a true author) has the minimal relative entropy.

Now we shall give a list of 82 authors included to the experiment. There are a lot of well-known Russian writers of the XIX and XX centuries among them: K. Bulychev, O. Avramenko, A. Bol'nykh, A. Volkov, G. Glazov, M. i S. Djachenko, A. Etoev, A. Kabakov, V. Kaplan, S. Kazmenko, V. Klimov, I. Krashevskij, I. Kublickaja, L. Kudrjavcev, A. Kurkov, Ju. Latynina, A. Lazarevich, A. Lazarchuk, S. Lem, N. Leonov, S. Loginov, E. Lukin, V. Chernjak, A.P. Chekhov, I. Khmelevskaja, L. and E. Lukiny, S. Luk'janenko, N.Ju. Markina, M. Naumova, S. Pavlov, B. Rajjnov, N. Rerikh, N. Romaneckijj, A. Romashov, V. Rybakov, K. Serafimov, I. Sergievskaja, S. Scheglov, A. Schegolev, V. Shinkarev, K. Sitnikov, S. Snegov, A. Stepanov, A. Stoljarov, R. Svetlov, A. Sviridov, E. Til'man, D. Truskinovskaja, A. Tjurin, V. Jugov, A. Molchanov, F.M. Dostoevskij, N.V. Gogol', D. Kharms, A. Zhitinskij, E. Khaeckaja, V. Khlumov, V. Kunin, A. Melikhov, V. Nabokov, JU. Nikitin, V. Segal', V. Jan, A. Tolstoj, I. Efremov, E. Fedorov, O. Grinevskij, N. Gumilev, L.N. Tolstoj, V. Mikhajlov, Ju. Nesterenko, A.S. Pushkin, L. Reznik, M.E. Saltykov-HHedrin, V. Shukshin, S.M. Solov'ev, A. Kac, E. Kozlovskij, S. Esenin, A. Strugackij, A. and B. Strugackie and B. Strugackij.

Note that transition frequencies of letter pairs are frequent, easily quantifiable and relatively immune from conscious control. The author of this study performed the same computations on English, French, and German texts and received similar results. Its description needs a separate paper and is beyond the scope of this short thesis. Clearly, further studies in this direction may be the best approach to "stylometry's 'holy grail', the fully automated identifier" (Holmes, 1998).

Holmes, D.I. (1998) The Evolution of Stylometry in Humanities Scholarship. Literary and Linguistic Computing, 13: 111–17.