

On an Application of Relative Entropy

Dmitry V. Khmelev^{1,*} and William J. Teahan^{2,†}

¹*Heriot-Watt University, Edinburgh, U.K. and Moscow State University, Russia*

²*University of Wales, Bangor, Dean Street, Bangor, LL57 1UT, U.K.*

(Dated: May 11, 2003)

We show that in problems of authorship attribution and other linguistic applications, a Markov Chains approach is a more attractive technique than Lempel-Ziv based compression.

PACS numbers: 89.70.+c, 01.20.+x, 05.20.-y, 05.45.Tp

We regret to point out several inaccurate and misleading statements that Benedetto *et al.* make in their Reply[1] to our Comment[2] on their paper titled “Language Trees and Zipping”[3].

First they confusingly state in paragraph 7 that Russian and Greek alphabets are not phonetic, putting Russian and Greek in a row with Chinese, the latter enjoying hieroglyphic writings. Second, they use unfair and irrelevant experiments in order to convince the reader that the gzip-based approach is better than the Markov chains based approach. Third, the figures reported for Newsgroups corpus seems to be obtained on a randomly selected small subset of the Newsgroups, which probably makes them completely meaningless in the discussed topic. Fourth, their reference to RAR compressor classification performance for refuting our Comment is irrelevant to our Comment and their Letter[3]. And fifth, authors of [1] obviously experience some problems with scientific English language. We elaborate on each of these points in more detail in the subsequent paragraphs.

It is a well-established fact that Russian language as well as Greek enjoys phonetic alphabet. Perhaps, Benedetto *et al.* [3] meant to use the transliteration for the construction of Language Tree (LT). However this procedure has its drawbacks like non-uniqueness, non-reversability, and inexactness of the transformation. Most importantly this procedure requires some *knowledge* about the language, which shows that the requirement for *a-priori* information, pointed out in [2] remains valid contrary to the claim in [3].

We believe that if one wants to compare the performance of several classification methods then the comparison should be performed in the same experimental framework. To start with, let us denote by M , G , g the classification performance of the following methods, respectively, on the corpus, discussed in [4]: Markov Chains approach [4]; attribution with a single source using gzip [5]; and attribution with multiple-source using gzip [3]. Let us denote by M' , G' , g' the classification performance of these methods in the framework of [3], and, finally, let M'' , G'' , g'' denote the classification performance of the same methods on the Newsgroups dataset. Notice that in [3] only values for $G'' = 60\% < g'' = 85\%$ and $G''' = 77\% < g''' = 93\%$ are

presented. One can not make any conclusion about M' or M'' using these data, so our statement about superiority of Markov Chains approach with respect to gzip approach (either with a single- or multiple-reference files) remains valid. Moreover, we have stated in our Comment [2] that $M = 69/82 \approx 84\%$ is greater than $G = 50/82 \approx 61\%$. We also reported to editors of Phys.Rev.Lett. in our answer to the referee report of Benedetto *et al.* that $g = 53/82 \approx 65\%$, which can indeed be considered as an argument for our claim that generally Markov chains are more attractive than gzip-based approach.

In our opinion the “slightly different method” of [3] should be considered as an approach to the design of the experiment, which leads to an extremely slow classification speed especially in the case of thousands of documents to classify, where a thousand source documents makes prohibitive the really large experiment on classification. This gives rise to the question of the validity of the figures $G'' = 60\%$ and $g'' = 85\%$ outlined in [1]. Traditionally, the precision of the classification method on the Newsgroups is measured in the following way: one performs a random 10-fold or 5-fold split and reports the average results of cross-validation. Typical numbers reported are around 80% [6], with 82.1% for PPM (Markov-based) approach. It would be interesting to know the technique used by[1], since even 5-fold split validation by their method would require about $5 \times (18828/5) \times (4 \times 18828/5) \approx 284 \times 10^6$ calls of gzip compression program, which is prohibitive on conventional computers. If one wants to apply a complete cross-validation as suggested in [3], then one has to do even more $18828^2 \approx 354 \times 10^6$ calls of gzip. We suspect that the figures $G'' = 60\%$ and $g'' = 85\%$, outlined in [1], are obtained on a randomly selected small subset of the Newsgroups, are subject to essential random variation, and hence $G'' = 60\%$ and $g'' = 85\%$ should not be used even for quantitative comparison of a single- and multiple-reference file setting.

Finally, in our Comment we stated that Markov chain approach as reported in [5] is superior to LZ approach *used in* [3]. This statement was misinterpreted in [1] as a general statement that LZ approach is outperformed by the *simple Markov chain* approach and Benedetto *et al.*[1] easily refute the misinterpreted statement using

our own result on RAR [5]. The correct generalization (and the only possible understanding in view of references given) of our statement is: *for any modification of LZ compression scheme there exists a modification of Markov Chain approach (PPM compression scheme), which outperforms LZ in classification* (this statement is similar to a well-known postulate among specialists: any modification of LZ compression scheme can be outperformed by a properly modified PPM compression scheme). The highly sophisticated, going far beyond the naive use of Ziv-Lempel theory, algorithm of RAR, know-how of its creator Eugene Roshal, should be compared with, for example, the the state-of-art PPMd (PPMonstr) algorithm developed recently by Dmitry Shkarin. And we find extremely interesting and scientifically valuable that the tough first-order Markov chain produce results competitive to highly sophisticated algorithms. As for the polemical comparison between Markov Chain and RAR compressor by Benedetto *et al.*[1], we find it irrelevant in the framework of their paper [3] as soon as they stand for technical details, like multiple- and single- source classification.

As a final remark we would like to point out that Reply [1] exhibits some language problems of it's authors themselves. Indeed, they reference to our comment [2] using expression "Khmelev *et al*" as if [2] has at least three co-authors (common meaning of *et al* is *and others*).

To sum up, one can not draw any conclusion on the comparison between n th order Markov chain approach (by which we meant[2] the PPM approach as well) with gzip-based approach from the statements, given in [1]. We also believe that the authors of[1] were not aware of our reported figure for $g = 64\%$; otherwise it looks very strange that they did not mention this argument in their Reply. Also we suggest to authors of [1] to present a fair comparison of their method against others, [6], [5] and, e.g., SVM approach [7–9].

P.S. This story shows that editors of physical journal like Phys.Rev.Lett. perhaps should avoid publishing papers like [3], because Phys.Rev.Lett. referees do not have enough experience to identify scientific value and mistakes in non-physical papers. We also encourage physicists and mathematicians to send their non-physical and non-mathematical papers to appropriate scientific journals, even if they are not so-well-known as Phys.Rev.Lett.

Probably such a publication would not yield much publicity, but the quality and scientific value of the paper would increase significantly.

The example with [3] is not unique. A similar story, which reappears time-to-time in newspapers, is the story about computing using DNA, described in details in [10]. It is possible to do computations with DNA. However, the amount of DNA, required for solution of, say, salesman problem with 100 cities, is comparable with the Earth mass, which makes it's use impractical and impossible. Notice that computer science methods allow to solve practical salesman problem in reasonable time for number of cities like 10^4 on contemporary computers. However, the authors of DNA computation speculative approach speculate that the effectiveness issue will be solved in future, a strange analogy with suggestion of [1].

Notice also that a publication of non-physical paper in physical journal evidence the crisis in physics, which responsible phisicists should aware. Otherwise why phisicists are publishing speculative papers on non-physical subjects? Is not this the evidence that they can not find application of their abilities in physics?

* Electronic address: D.Khmelev@newton.cam.ac.uk

† Electronic address: wjt@informatics.bangor.ac.uk

- [1] D. Benedetto, E. Caglioti, and V. Loreto, Phys. Rev. Lett. **90** (2003).
- [2] D. Khmelev and W. Teahan, Phys. Rev. Lett. **90** (2003).
- [3] D. Benedetto, E. Caglioti, and V. Loreto, Phys. Rev. Lett. **88** (2002).
- [4] D. Khmelev, J. of Quantitative Linguistics **7**, 201 (2000).
- [5] O. Kukushkina, A. Polikarpov, and D. Khmelev, Problems of Information Transmission **37**, 172 (2001).
- [6] W. J. Teahan, Proceedings RIAO'2000 **2**, 943 (2000).
- [7] E. L. Allwein, R. E. Schapire, and Y. Singer, in *Proc. 17th International Conf. on Machine Learning* (Morgan Kaufmann, San Francisco, CA, 2000), pp. 9–16, URL citeseer.nj.nec.com/allwein00reducing.html.
- [8] K. Crammer and Y. Singer, in *Computational Learning Theory* (2000), pp. 35–46, URL citeseer.nj.nec.com/crammer00learnability.html.
- [9] J. Weston and C. Watkins, *Multi-class support vector machines* (1998), URL citeseer.nj.nec.com/8884.html.
- [10] D. Gusfield, *Algorithms on strings, trees, and sequences* (Cambridge University Press, Cambridge, 1997).