Автоматическая периодизация авторских корпусов

Балуева Дарья Владимировна, РГГУ

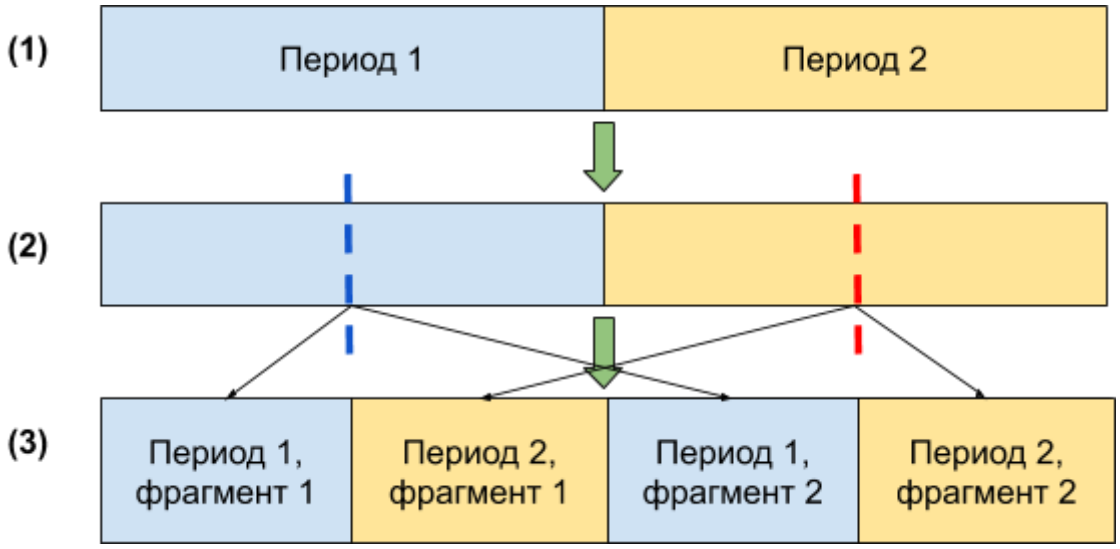
Цель нашей работы — предложить и проверить автоматический метод периодизации творчества. Мы предполагаем, что если упорядочить все тексты какого-либо автора по дате создания, то внутри них можно провести одну или несколько границ, разделяющих тексты на максимально непохожие между собой группы. Выделенные группы можно будет условно назвать ранним и поздним творчеством (если их две), ранним, зрелым и поздним творчеством (если их три) и т. п.

Подобные исследования уже проводились ранее: например, в [Reeve 2018; Salgaro, Rebora 2018] на материале художественной прозы проверяется возможность автоматического выделения «раннего» и «позднего» стиля писателей, а в [Dobson 2019] предлагался метод автоматической периодизации исторических текстов. В существующих работах границы и количество периодов определяются самими исследователями, мы же проверим, можно ли выделять разные периоды автоматически.

Наш материал — корпуса 75 авторов из поэтического подкорпуса НКРЯ, объем которых более чем 50 тыс. словоформ. Тексты каждого автора, размеченные и отсортированные по дате создания, мы автоматически делим на группы всеми возможными способами. При каждом разбиении из текстов обеих групп создаются частотные списки символьных *n*-грамм длины 4, и между ними вычисляется косинусное расстояние. Те разбиения, при которых расстояние оказывается наибольшим, мы считаем отражающими разные периоды творчества.

Эффективность косинусной меры и символьных n-грамм для измерения сходства текстов и автоматического определения авторства уже подтверждалась [Juola 2006; Houvardas, Stamatatos 2006; Evert et al. 2017], в том числе — на материале, аналогичном материалу данной работы [Piperski 2019]. Разбив все корпуса на части, мы проанализируем, насколько четко выделяются разные периоды. Наша задача — установить, от каких свойств корпусов и отдельных текстов зависит разбиение, на сколько периодов мы можем успешно делить творчество авторов. При необходимости мы подберем более подходящую меру расстояния или другие языковые уровни для сравнения. [Rayson et al. 2000; Burrows 2002; Gomaa, Fahmy 2013]

В конце мы проверим устойчивость границ дополнительным разбиением и смешиванием выделенных групп. Например, вот так:



Корпуса будут разделены на части-периоды, максимально непохожие друг на друга **(1)**. Каждый из этих периодов мы вручную разобьем еще раз **(2)**. Полученные фрагменты перемешаем и соединим так, чтобы границы разных фрагментов одного периода не накладывались **(3)**. Затем мы попробуем отделить эти четыре фрагмента друг от друга автоматически, снова найдя расстояния между ними. В конце мы проверим, насколько четко фрагменты разных периодов отделятся друг от друга.

Итак, мы надеемся получить метод, который позволит делить творчество отдельного автора (корпус текстов, размеченных по дате создания) на периоды. Этот метод может быть применим для изучения развития авторского стиля — как авторов, представленных в корпусе, так и других.

Литература

1. Argamon S. (2008). Interpreting Burrows' delta: Geometric and probabilistic foundations. Literary and Linguistic Computing , 23(2): 131–47.
2. Burrows J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. Literary and linguistic computing, 17(3), 267-287.
3. Dobson James E. (2019) Critical Digital Humanities: The Search for a Methodology. University of Illinois Press.
4. Eder M. (2015). Taking stylometry to the limits: Benchmark study on 5,281 texts from Patrologia Latina. In Digital Humanities 2015: Conference Abstracts.
5. Evert S., Proisl, T., Jannidis F., Reger I., Pielström S.,Schöch C., & Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution, Digital Scholarship in the Humanities, Volume 32, Issue suppl_2, December 2017, Pages ii4–ii16.
6. Gomaa W. H., Fahmy A. (2013). A Survey of Text Similarity Approaches International / Journal of Computer Applications, 68(13), April, pp. 13–18.

7. Holmes D.I. (1994). Authorship attribution. Computers and the Humanities, 28(2), 87– 106. https://doi.org/10.1007/BF01830689
8. Hoover D. (2004). Delta prime? Literary and Linguistic Computing , 19(4): 477–95.
9. Houvardas J., & Stamatatos E. (2006). N-Gram Feature Selection for Authorship Identification. In J. Euzenat & J. Domingue (Eds.), Artificial Intelligence: Methodology,
10. Juola P. (2006). Authorship attribution. Foundations and Trends in Information Retrieval , 1(3): 233–334.
11. Kilgarriff A. (2001). Comparing corpora. International Journal of Corpus Linguistics, 6:1, 97–133.
12. Kjell B., Woods W.A., & Frieder O. (1994). Discrimination of authorship using visualization. Information Processing & Management, 30(1), 141–150. https://doi.org/10.1016/0306-4573(94)90029-9
13. Piperski A. (2019). Authorship Attribution with a Very Naïve Bayes Model and What It Can Tell Us about Russian Poetry. In Computational Linguistics and Intellectual Technologies. Papers from the Annual International. Conference “Dialogue”. Issue 18
14. Rayson P., & Garside R. (2000). Comparing corpora using frequency profiling. / In Proceedings of the Comparing Corpora Workshop at ACL 2000. Hong Kong.
15. Reeve J. P. (2018). “Does ‘Late Style’ Exist? New Stylometric Approaches to Variation in Single-Author Corpora”, in DH2018 Book of Abstracts, ADHO, Mexico City, pp. 478-481.
16. Salgaro M., & Rebora S. (2018). Is “Late Style” measurable? A stylometric analysis of Johann Wolfgang Goethe’s, Robert Musil’s, and Franz Kafka’s late works. Elephant&Castle, Systems, and Applications (pp. 77–86). Berlin, Heidelberg: Springer.