- увеличение ширины контекста приводит к улучшению качества, но только до определённого момента (использование тетраграмм дало заметное ухудшение результатов). Отметим, что применение популярной концепции word2vec [19] дало очень слабый результат (0.83–0.85), и при этом время подсчёта увеличилось в несколько раз;
- на основе лексических признаков или *n*-грамм можно получить хороший результат даже с помощью простых классификаторов;
- исходя из критерия максимизации минимальной точности, следует использовать многослойный персептрон, а в качестве лексических характеристик стихотворений – триграммы.

5 Заключение

В работе проанализированы принципы формирования обучающих выборок для алгоритмов определения стилей и жанровых типов. Проведены вычислительные эксперименты с использованием корпуса текстов лицейской лирики А.С. Пушкина по выбору наиболее точного алгоритма классификации поэтических текстов, в том числе с использованием наиболее известных приемов ансамблирования базовых алгоритмов в композиции, таких, как взвешенное голосование, бустинг и стекинг, причем в качестве характеристических признаков стихотворений использовались одиночные слов, биграммы и триграммы. Рассмотренные алгоритмы показали свою работоспособность (при этом, исходя из критерия максимизации минимальной точности, следует использовать многослойный персептрон, а в качестве лексических характеристик стихотворений – триграммы) и могут быть использованы для автоматизации комплексного анализа русских поэтических текстов, существенно облегчая работу эксперта при определении их стилей и жанров путем предоставления соответствующих рекомендаций.

Поддержка

Работа выполнена при частичной поддержке Президиума РАН (проект 2016-PRAS-0015) и Президентской программы «Ведущие научные школы РФ» (грант 7214.2016.9).

Литература

[1] Барахнин, В.Б., Кожемякина, О.Ю. Об автоматизации комплексного анализа русского поэтического текста. CEUR Workshop Proceedings, 934, сс. 167-171 (2012)

[2] Лесцова, М.А.: Определение ядра и периферии жанров оды, песни, послания, элегии и эпитафии на материале английских поэтов-сентименталистов XVIII века. Вестник Челябинского государственного пед. университета, 4, сс. 196-205 (2014)

[3] Орлов, Ю.Н., Осминин, К.П.: Определение жанра и автора литературного произведения статистическими методами. Прикладная информатика, 26 (2), сс. 95-108 (2010)

[4] Орлов, Ю.Н., Осминин, К.П.: Методы статистического анализа литературных текстов. Эдиториал УРСС, Москва (2012)

[5] Barakhnin, V., Kozhemyakina, O., Pastushkov, I.: Automated Determination of the Type of Genre and Stylistic Coloring of Russian Texts. ITM Web of Conferences 10, 02001, 4 p. (2017). doi: 10.1051/itmconf/20171002001

[6] Vapnik,V.N.: The Nature of Statistical Learning Theory. Springer-Verlag (1995)

[7] Грехнев, В.А.: Лирика Пушкина. О поэтике жанров. Горький: Волго-Вятское книжное издательство (1985)

[8] Барахнин, В.Б., Кожемякина, О.Ю.: К проблеме аутентичности фонетического анализа в связи с возможными особенностями авторской орфографии (на примере чередования окончаний -ой/-ый в лирике А.С. Пушкина). Вестник Томского государственного университета. Филология, 13 (2), сс. 5-28 (2016)

[9] Ломоносов, М.В.: Предисловие о пользе книг церковных в российском языке. Ломоносов, М. В. Полн. собр. соч. 7, сс. 585-592. М.–Л.: Изд-во АН СССР (1952)

[10] Магомедова, Д.М.: Филологический анализ лирического стихотворения. М.: Издательский центр «Академия» (2004)

[11] Свободина, С.Ф.: К вопросу о философской направленности и жанровых особенностях стихотворения А.С. Пушкина «Безверие». Пушкинский музеум: альманах, 6, сс. 261-270. Всероссийский музей А.С. Пушкина, Санкт-Петербург (2014)

[12] Ахманова, О.С.: Очерки по общей и русской лексикологии. М.: Учпедгиз (1957)

[13] Freund, Y., Schapire, R.E.: A Short Introduction to Boosting. J. of Japanese Society for Artificial Intelligence, 14 (5), pp. 771-780 (1999)

[14] Friedman, J.H.: Stochastic Gradient Boosting. Computational Statistics and Data Analysis, 38 (4), pp. 367-378 (2002)

[15] Wolpert, D.H.: Stacked Generalization. Neural Networks, 5 (2), pp. 241-259 (1992)

[16] Chawla, N.V.: Data Mining for Imbalanced Datasets: An Overview. Data Mining and Knowledge Discovery Handbook, pp. 875-886. Springer-Verlag (2010)

[17] Cover, T.M., Hart, P.E.: Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, 13, pp. 21-27 (1967)

[18] Korobov, M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science, 542, pp. 320-332 (2015)

[19] Mikolov, T., Kai, Chen, Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. Computation and Language, 12 p., (2013). https://arxiv.org/pdf/1301.3781.pdf