УДК 82-1 + 81'322
DOI 10.25205/1818-7935-2017-15-3-13-23

В. Б. Барахнин [1, 2], **О. Ю. Кожемякина** [1]
И. С. Пастушков [1], **Е. В. Рычкова** [1, 2]

[1] *Институт вычислительных технологий СО РАН*
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия

[2] *Новосибирский государственный университет*
ул. Пирогова, 1, Новосибирск, 630090, Россия

bar@ict.nsc.ru, olgakozhemyakina@mail.ru
pas2shkov.ilya@gmail.com, helen@ict.nsc.ru

АВТОМАТИЗИРОВАННАЯ КЛАССИФИКАЦИЯ РУССКИХ ПОЭТИЧЕСКИХ ТЕКСТОВ ПО ЖАНРАМ И СТИЛЯМ *

Проанализированы принципы формирования обучающих выборок для алгоритмов автоматизированного определения жанров и стилей русских поэтических текстов. Проведены вычислительные эксперименты с использованием корпуса текстов лицейской лирики А. С. Пушкина по выбору наиболее точного алгоритма классификации, 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 слова, биграммы и триграммы. Рассмотренные алгоритмы показали свою работоспособность и могут быть использованы для автоматизации комплексного анализа русских поэтических текстов, при этом было установлено, что даже с помощью простых классификаторов на основе лексических признаков или *n*-грамм можно получить хороший результат; исходя из критерия максимизации минимальной точности, следует использовать многослойный персептрон, а в качестве лексических характеристик стихотворений – триграммы. Разработанные алгоритмы способны существенно облегчить работу эксперта при определении стилей и жанров поэтических текстов путем предоставления соответствующих рекомендаций.

Ключевые слова: автоматический анализ поэтических текстов, определение жанров и стилей, алгоритмы классификации.

В задачах автоматизированного анализа текстов на естественном языке возникает проблема определения их жанрового типа и стилистических характеристик. С этой проблемой исследователь может столкнуться в широком спектре ситуаций: от задач автоматизации комплексного анализа поэтических текстов, для которых жанровый тип и стилистические характеристики являются важными атрибутами, используемыми при определении влияния низших уровней стиха на высшие (см., например, [Барахнин, Кожемякина, 2012]), до отслеживания сообщений в социальных сетях с целью выявления террористических угроз, определения маркетинговых предпочтений покупателей и т. п.

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

Барахнин В. Б., *Кожемякина О. Ю.*, *Пастушков И. С.*, *Рычкова Е. В.* Автоматизированная классификация русских поэтических текстов по жанрам и стилям // Вестн. Новосиб. гос. ун-та. Серия: Лингвистика и межкультурная коммуникация. 2017. Т. 15, № 3. С. 13–23.

ISSN 1818-7935
Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2017. Том 15, № 3

Исследования в области автоматизированного определения жанрового типа текстов начаты недавно – в начале 2010-х гг. Так, в работе [Лесцова, 2014] предложены алгоритмы определения жанров оды, песни, послания, элегии и эпитафии на материале английских поэтов-сентименталистов XVIII в. Временной период в этом исследовании был выбран не случайно: в поэзии XVIII в. господствовал классицизм с его жесткими жанровыми канонами, что существенно облегчило разработку алгоритмов.

В статье [Орлов, Осминин, 2010] изложен метод классификации текстов (по определенным жанрам и по авторам) на основе анализа статистических закономерностей буквенных распределений, т. е. вероятностей встречаемости букв и буквосочетаний, при этом подчеркнуто, что решение найдено без «вторжения в область литературы, т. е. без анализа синтаксиса, литературных приемов и схем взаимодействий персонажей». Однако в работе [Орлов, Осминин, 2012] сами авторы строят оригинальный контрпример к статистическому методу идентификации, что, в свою очередь, показывает необходимость использования хотя бы методов морфологического анализа.

Что же касается автоматизации определения стилистических характеристик текстов, то нам неизвестны исследования в этой области по крайней мере для текстов на русском языке.

В работе [Barakhnin et al., 2017] показано, что метод опорных векторов (support vector machine, SVM) [Vapnik, 1995] позволил получить хорошие результаты при определении стилей поэтических текстов и удовлетворительные – при определении жанров.

В настоящей работе мы расширили используемые подходы, в частности, учитывая при построении характеристического вектора используемых в стихотворении лексем количество их вхождений, а также проводя эксперименты с характеристическими векторами биграмм и триграмм. Кроме того, нами был проведен сравнительный анализ целого ряда алгоритмов классификации поэтических текстов, в том числе с использованием наиболее известных приемов ансамблирования, т. е. построения композиций алгоритмов, в которых ошибки отдельных алгоритмов взаимно компенсируются. При ансамблировании рассматриваются алгоритмы, в которых функция, называемая алгоритмическим оператором, устанавливает соответствие между множеством объектов и пространством оценок, а функция, называемая решающим правилом, устанавливает соответствие между пространством оценок и множеством значений целевой функции. Таким образом, рассматриваемые алгоритмы имеют вид суперпозиции алгоритмического оператора и решающего правила. Многие алгоритмы классификации имеют именно такую структуру: сначала вычисляются оценки принадлежности объекта классам, затем решающее правило переводит эти оценки в номер класса. Значением оценки может быть вероятность принадлежности объекта классу, расстояние от объекта до разделяющей поверхности, степень уверенности классификации и т. п.

Таким образом, в представленной статье проведен сравнительный анализ целого ряда методов автоматизированной классификации поэтических текстов, включая наиболее известные приемы ансамблирования базовых алгоритмов в композиции: взвешенное голосование, бустинг и стекинг.

1. *Построение совместного классификатора жанровых типов и стилей поэтических текстов*. При построении совместного («двумерного») классификатора жанровых типов и стилей текстов мы учитывали, что классификатор как таковой представляет собой многомерную структуру, основанную на совокупности определяющих объект исследования параметров. Многомерность структуры уже сама по себе предполагает определенную погрешность в конечном результате анализа, поскольку чем больше мы вводим в систему данных, тем больше потенциальная вероятность появления многовариативности на каждом из этапов анализа. Поэтому при построении многомерных классификаторов, связанных с такими сложными для однозначного определения категориями, как жанр и стиль, необходима поэтапная разработка каждого параметра анализа с целью исключения возможных погрешностей и вариативности результата, что в случае анализа неструктурированных текстов приобретает решающее значение.

Нами разрабатывается совместный («двумерный») классификатор жанровых типов и стилей текстов. Такой классификатор создается впервые (по крайней мере для текстов на русском языке).

Создание «универсального» классификатора художественных текстов по их жанрово-стилистическим характеристикам требует обобщения огромного эмпирического материала, поэтому мы решили остановиться на классификации лицейской лирики А. С. Пушкина, поскольку в классической теории жанр произведения строго диктует выбор того или иного стиля (применительно к стихотворным текстам эта зависимость подробно рассмотрена в монографии Д. М. Магомедовой [2004]). Обоснования такого выбора приведены ниже.

Наиболее эффективным подходом к автоматизации определения жанровых типов является использование алгоритмов с обучением. Однако формирование обучающей выборки – задача отнюдь не банальная. Наша попытка использовать в качестве обучающей выборки пушкинскую лирику зрелого периода (1828–1831 гг.) потерпела неудачу уже на раннем этапе работы, поскольку жанровое разнообразие пушкинского творчества этого периода в соотношении со стилевыми особенностями произведений в особой пушкинской манере не подчиняется общепринятым законам. На данную черту указывал ранее В. А. Грехнев: «Жанры и стиль не противостоят друг другу как враждебные, взаимоотрицающие начала, но между ними всегда существует внутреннее напряжение. Напряжение это возрастает там, где возрастают мощь и размах писательской индивидуальности» [1985. С. 234]. Отсюда возникают жанрово-стилистические разновидности и варианты, во «внутреннем напряжении» между стилем и жанром берут начало неканонические жанры [Магомедова, 2004], и именно для обучающей выборки это становится критичным, поскольку возникают особенности, не попадающие в систему и, следовательно, противоречащие по своей сути материалу для построения жанрово-стилевой системы. Вследствие этого мы решили остановиться на лицейской лирике, поскольку в ней наблюдается использование наиболее строгих жанровых форм, стилистическое единство, а также следование правилам грамматики своего времени: «Почти вся лицейская лирика относится к возвышенному стилю, исключение – всего несколько стихотворений. Даже многие сатирические стихи написаны вполне в возвышенном стиле. Можно утверждать, что в ранних стихах Пушкина чувствуется влияние жестких правил “Грамматики” его лицейского учителя Н. Ф. Кошанского» [Барахнин, Кожемякина, 2016. С. 24].

В свою очередь, использование именно лицейской лирики как материала для создания обучающей выборки оправдано и стилевым аспектом, поскольку стилевая дифференциация лексем – этап разработки классификатора. Для текстов на русском языке принято восходящее к трудам М. В. Ломоносова [1952] деление текстов (прежде всего, художественных) на относящиеся к высокому, среднему и низкому стилям. Исторически каждый из них характеризуется специфическим соотношением использования старославянских (церковнославянских) и собственно русских слов (при этом отдельно рассматривается группа слов, общих для старославянского и русского языков), долей архаизмов, а также употреблением определенных синтаксических конструкций.

Для реализации поставленной задачи мы идем от практики, делая выборку произведений Пушкина лицейского периода как материала, на котором вероятно построение наиболее точной теоретической модели жанрово-стилистических зависимостей, что делает конечный результат анализа наиболее точным и позволяет разработать наиболее адекватный классификатор, относящийся к стилевому аспекту. Поскольку мы решили ограничиться анализом жанров только малых стихотворных форм, то из анализа исключены поэмы, сказки, переводы, Dubia. Далее делаем список, включающий в себя стихотворения, как соответствующие системе жанров, приведенной в работе [Там же], так и не входящие в эту систему.

В итоге рассмотрения списка произведений, взятого нами для анализа, мы выделяем следующие группы жанров.

Канонические: ода – 4 произведения; элегия – 27 произведений (в том числе одна историческая элегия – «Наполеон на Эльбе»); идиллия – 2; послание – 55; баллада – 3; неканонических произведений, выделенных Д. М. Магомедовой (фрагмент, рассказ в стихах), нет.

Также мы добавляем жанры, которых нет в разработанной Д. М. Магомедовой системе канонических – неканонических: эпиграмма – 18 произведений; мадригал – 4; сонет – 1; романс – 1; анекдот – 1; притча – 2. Кроме этого, стихотворение «Безверие» (1817) определяется как элегия и философская ода [Свободина, 2014], но для анализа мы определяем его как философскую оду.

Таблица 1

Статистика по жанрово-стилевому соответствию

Жанр	Стиль		
	высокий	средний	низкий
Ода	4	–	–
Притча	1	1	–
Мадригал	4	–	–
Послание	–	55	5
Идиллия	–	2	–
Элегия	–	37	–
Романс	–	1	–
Баллада	–	3	–
Эпиграмма	–	–	18
Анекдот	–	–	1

Жанровые типы этих произведений легли в основу классификатора (табл. 1): по одной оси мы разместили жанровые типы в порядке возрастания «возвышенности»: ода, элегия, идиллия, послание и т. д., а по другой оси – традиционные стили.

На данном эмпирическом материале просматривается очевидная корреляция между жанровыми и стилистическими характеристиками текстов: ода, элегия и идиллия обычно написаны высоким стилем, в них не используется лексика, соответствующая низкому стилю, а для эпиграмм, напротив, характерно использование элементов лексики низкого стиля. Вообще говоря, стиль текста определяется по наиболее «низким» его лексемам, что особенно характерно для эпиграмм: наличие высокой лексики, употребляемой нередко в ироническом ключе, не должно вводить в заблуждение, ибо употребление одного-двух слов разговорной или откровенно обсценной лексики сразу характеризует авторский замысел. Тем не менее для жанров, традиционно предполагающих возвышенную форму, прежде всего мадригала, мы не считаем целесообразным относить принадлежащие к ним стихотворения, в которых с ироническими целями употреблено несколько «сниженных» (но не обсценных!) слов, к сниженному стилю. Следует отметить, что специфика стиля проявляется на лексическом уровне в гораздо большей степени, чем жанр.

В нашей выборке в силу ее специфических задач произведения, написанные в жанре притчи, отнесены: одно («Наездники») к высокому стилю, второе («Истина») к среднему, хотя, как известно, притча, будучи жанром наиболее близким к басне, предполагает возможность написания ее в разных стилях, о чем свидетельствует, в частности, притча Пушкина «Сапожник», которую можно отнести, скорее, к низкому («разговорному») стилю.

2. *О возможности создания словаря стилистически дифференцированных лексем*. Прежде чем приступить к выбору алгоритмов определения стилистических и жанровых характеристик поэтических текстов, необходимо решить вопрос: возможно ли использовать для решения этой задачи априори составленные словари лексем, имеющих ту или иную стилистическую или жанровую окраску?

Большое внимание вопросам стилистической дифференциации слов уделено в монографии О. С. Ахмановой «Очерки по общей и русской лексикологии» [1957]. Приведены списки слов «разговорных», со «сниженной» стилевой характеристикой и с «повышенной» стилевой характеристикой. Однако эти списки далеко не полны и носят, скорее, иллюстративный характер, более того, автор признает, что «далеко не все из включенных в них слов будут одинаково убедительными (многие, несомненно, покажутся спорными)», и, наконец, стилистическая окраска некоторых лексем менялась со временем, т. е. эта характеристика, взятая из монографии [Там же], могла быть иной как для языка XIX в., так и для современного. Поэтому для соотнесения слова с тем или иным стилем в той же монографии предлагается ис-

пользовать анализ их структурно-семантической формы. Так, существительные с суффиксом *-к-* в разнообразных структурно-семантических вариантах, а также с различными суффиксами со значением «лица» относятся к «разговорной» или «сниженной» лексике; для «разговорной», в отличие от «сниженной», лексики характерно большое число наречий; для «книжной» лексики характерны заимствованные слова, а для «возвышенной» – славянские со сложной структурой, а также архаизмы и т. п.

Однако все эти наблюдения носят весьма частный характер. Так, слова с суффиксом *-к-* *пытка*, *речка*, *шутка* и др. встречаются в стихах Пушкина, относящихся отнюдь не к «низкому» или «разговорному» стилю, то же самое относится к словам *бочка*, *кружка*, *пушка* и др., в которых *-к-* является частью корня, но установление этого факта требует нетривиального этимологического анализа, плохо поддающегося автоматизации. Заимствованные слова с течением времени становятся достоянием всех стилей, и это касается не только «древних» заимствований вроде *лошадь* или *собака*, но и новых: *велосипед*, *танк* и т. п. Славянизмы, в том числе со сложной структурой, могли использоваться, в том числе, для придания стихотворению иронического оттенка (например, «Ода его сиятельству графу Д. И. Хвостову» Пушкина и многочисленные сатирические стихи А. К. Толстого).

Ситуация осложняется еще и тем, что нередко «разговорным» или «сниженным» является не все слово, а лишь один из его лексико-семантических вариантов, а также обретением словом той или иной окраски лишь при вхождении в состав фразеологизма.

Таким образом, вхождение в текст отдельных лексем не может служить достаточно надежным критерием отнесения текста к определенному стилистическому типу. Тем более, четкое выделение жанровой принадлежности отдельных слов представляется совершенно бесперспективной задачей, и нам неизвестны сколько-нибудь удовлетворительные попытки ее разрешения хотя бы на теоретическом уровне. Именно поэтому нам представляется наиболее целесообразным определять стилистические и жанровые характеристики поэтических текстов на основании вхождения в них совокупности лексем, определяемой на базе обучающей выборки.

3. *Описание численного эксперимента*. Для эксперимента использовался описанный выше корпус текстов лицейской лирики Пушкина, состоящий из 121 стихотворения, размеченных экспертом по жанрам и стилям.

При обучении была проведена лемматизация всех уникальных слов, встречающихся в текстах, и создан словарь их исходных форм. Отдельно был составлен словарь имен собственных, которые удалялись из словаря всех слов, поскольку гипотезы, подобные той, что имена из древнегреческого пантеона присущи только высокому стилю, были опровергнуты, в частности, при подготовке данных для экспериментов. Каждый текст кодировался последовательностью цифр, соответствующей количеству вхождений в него слов из словаря: 0 ставился, если слова нет в тексте; 1 – если слово встречается 1 раз; 2 – если 2 раза и т. д. Помимо лексических признаков, первоначально предполагалось использовать стихотворные характеристики (рифма, размер, стопность и т. п.), но это привело к серьезному ухудшению качества классификации, поэтому было решено от них отказаться.

Также были собраны словари n-грамм ($n = 2, 3$), которые не содержали имен собственных, причем n-граммы были не упорядоченными внутри себя, поскольку в поэзии очень часто встречается обратный порядок слов.

Далее опишем применявшиеся нами приемы ансамблирования, т. е. комбинирования алгоритмов, взаимно улучшающего их свойства.

Во-первых, это два варианта взвешенного голосования с использованием нескольких классификаторов: в случае hard-голосования решение о классификации того или иного объекта принимается на основании заключения большинства используемых классификаторов; в случае soft-голосования результат определяется исходя из аргумента максимизации вероятности отнесения классифицируемого объекта к некоторому классу.

Во-вторых, это бустинг, идея которого состоит в применении принципа «жадного» выбора очередного алгоритма для добавления в композицию так, чтобы он лучшим образом компенсировал имеющиеся на этом шаге ошибки. Нами были применены наиболее известные примеры бустинга – AdaBoost [Freund, Schapire, 1999] и градиентный бустинг (Gradient

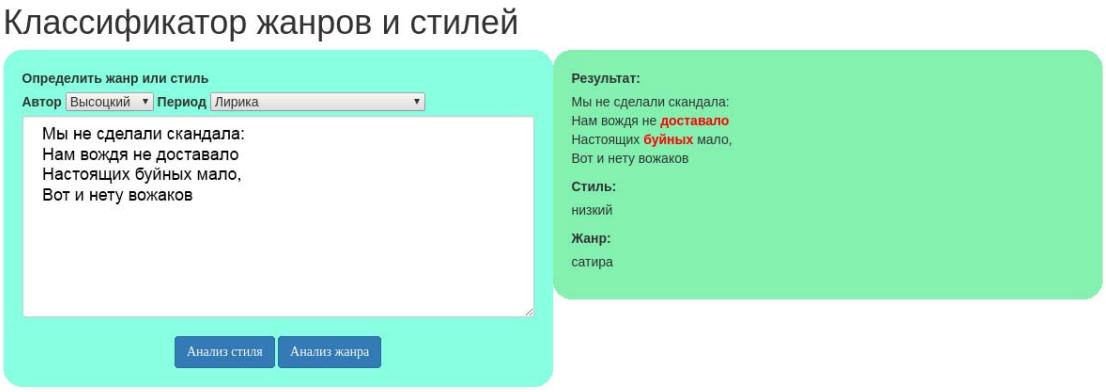
boosting) [Friedman, 2002]. Среди прочих, нами был применен метод опорных векторов (SVM) [Vapnik, 1995], усиленный AdaBoost.

Наконец, в-третьих, стекинг [Wolpert, 1992], который основан на применении базовых классификаторов для получения предсказаний (метапризнаков) и использовании их как признаков низшего ранга для некоторого «обобщающего» алгоритма (метаалгоритма). Иными словами, основной идеей стекинга является преобразование исходного пространства признаков задачи в новое пространство, точками которого являются предсказания базовых алгоритмов. В данном исследовании в качестве метаалгоритма была взята логистическая регрессия над SVM, градиентным бустингом, многослойным персептроном и голосованиями.

Отметим, что в процессе решения рассматриваемой задачи нам пришлось столкнуться с проблемой миноритарных классов, которые ясно обозначены в табл. 1. Для решения этой проблемы были применены случайное дублирование элементов миноритарных классов, а также стратегия SMOTE [Chawla, 2010], которая основана на идее генерации некоторого количества искусственных примеров, которые были бы «похожи» на имеющиеся в миноритарном классе, но при этом не дублировали их. Для создания новой записи вычисляют разность $u = X_b - X_a$, где X_a, X_b – векторы признаков «соседних» примеров a и b из миноритарного класса, которые находят, используя алгоритм ближайшего соседа [Cover, Hart, 1967]. Далее из u путем умножения каждого его элемента на случайное число в интервале (0, 1) получают $\hat{u}$. Вектор признаков нового примера вычисляется путем сложения X_a и $\hat{u}$. Алгоритм SMOTE позволяет задавать количество записей, которое необходимо искусственно сгенерировать. Степень сходства примеров a и b можно регулировать путем изменения значения k (числа ближайших соседей).

Программное приложение для классификации поэтических текстов реализовано на языке Python с использованием библиотек sklearn (реализация алгоритмов, их композиций и кроссвалидации), imblearn (реализация SMOTE), xgboost (наиболее эффективная реализация градиентного бустинга) и pymorphy2 [Korobov, 2015] для приведения слов к нормализованному виду, а также для отсечения имен собственных.

Численный эксперимент осуществляется с помощью интерфейса, в окно которого вносится исследуемый текст. Затем, после нажатия на кнопки «Анализ стиля» и «Анализ жанра», в правой части отображается результат классификации по стилю и жанру:



В табл. 2–7 приведены результаты работы классификаторов и их композиций, полученные при трехэтапной кроссвалидации (трехкратное разбиение корпуса на обучающее и тестовое множество, каждый раз классификатор обучался на обучаемом и оценивался на тестовом множестве). Из таблицы результатов был исключен рекомендуемый при работе со SMOTE метод ближайших соседей, так как он показывал очень низкую точность.

Таблица 2

Лексические признаки + SMOTE для определения стиля

Классификатор	Среднее	Max	Min
SVM AdaBoost	0,88	0,91	0,84
XGBoost	0,83	0,9	0,81
Многослойный персептрон	0,85	0,95	0,67
Голосование, hard	0,94	0,95	0,92
Голосование, soft	0,94	0,95	0,92
Стекинг	0,94	0,97	0,92

Таблица 3

Лексические признаки + случайное дублирование миноритарных классов для определения жанра

Классификатор	Среднее	Max	Min
SVM AdaBoost	0,88	0,89	0,86
XGBoost	0,90	0,92	0,89
Многослойный персептрон	0,93	0,95	0,91
Голосование, hard	0,92	0,95	0,88
Голосование, soft	0,92	0,96	0,88
Стекинг	0,90	0,93	0,87

Таблица 4

Биграммы + SMOTE для определения стиля

Классификатор	Среднее	Max	Min
SVM AdaBoost	0,98	1,00	0,96
XGBoost	0,92	0,94	0,90
Многослойный персептрон	0,98	1,00	0,95
Голосование, hard	0,98	1,00	0,96
Голосование, soft	0,97	0,98	0,95
Стекинг	0,98	0,99	0,95

Таблица 5

Биграммы + случайное дублирование миноритарных классов для определения жанра

Классификатор	Среднее	Max	Min
SVM AdaBoost	0,94	0,96	0,90
XGBoost	0,97	1,00	0,93
Многослойный персептрон	0,97	0,99	0,94
Голосование, hard	0,94	1,00	0,88
Голосование, soft	0,93	1,00	0,88
Стекинг	0,96	1,00	0,89

Таблица 6

Триграммы + SMOTE для определения стиля

Классификатор	Среднее	Max	Min
SVM AdaBoost	0,98	0,99	0,98
XGBoost	0,93	0,94	0,92
Многослойный персептрон	0,99	0,99	0,99
Голосование, hard	0,98	0,99	0,98
Голосование, soft	0,98	0,99	0,98
Стекинг	0,98	0,99	0,99

Таблица 7

Триграммы + случайное дублирование миноритарных классов для определения жанра

Классификатор	Среднее	Max	Min
SVM AdaBoost	0,95	1,00	0,86
XGBoost	0,94	1,00	0,84
Многослойный персептрон	0,97	0,99	0,95
Голосование, hard	0,96	1,00	0,91
Голосование, soft	0,96	1,00	0,91
Стекинг	0,96	1,00	0,88

Из полученных данных можно сделать следующие выводы:

- стекинг не всегда дает наилучшее (т. е. наиболее соответствующее экспертной оценке) решение (см. табл. 3);
- при увеличении контекста признаков (от одного слова к би- и триграммам) XGBoost становится более точным, чем многослойный персептрон;
- увеличение ширины контекста приводит к улучшению качества, но только до определенного момента (использование тетраграмм дало заметное ухудшение результатов, соответствующие таблицы из соображений экономии не приводятся). Отметим, что применение популярной концепции word2vec [Mikolov et al., 2013] дало очень слабый результат (0,83–0,85), и при этом время подсчета 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.

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 слова, биграммы и триграммы. Рассмотренные алгоритмы показали свою работоспособность и могут быть использованы для автоматизации комплексного анализа русских поэтических текстов.

Список литературы

Ахманова О. С. Очерки по общей и русской лексикологии. Моногр. М.: Учпедгиз, 1957.

Барахнин В. Б., *Кожемякина О. Ю.* Об автоматизации комплексного анализа русского поэтического текста // CEUR Workshop Proceedings. 2012. V. 934. P. 167–171.

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) // Вестн. Том. гос. ун-та. Серия: Филология. 2016. Т. 13, № 2. С. 5–28.

Грехнев В. А. Лирика Пушкина. О поэтике жанров: Моногр. Горький: Волго-Вят. кн. изд-во, 1985.

Лесцова М. А. Определение ядра и периферии жанров оды, песни, послания, элегии и эпитафии на материале английских поэтов-сентименталистов XVIII века // Вестн. Челяб. гос. пед. ун-та. 2014. Вып. 4. С. 196–205.

Ломоносов М. В. Предисловие о пользе книг церковных в российском языке // Ломоносов М. В. Полн. собр. соч. М.; Л.: Изд-во АН СССР, 1952. Т. 7. С. 585–592.

Магомедова Д. М. Филологический анализ лирического стихотворения: Моногр. М.: Академия, 2004.

Орлов Ю. Н., *Осминин К. П.* Методы статистического анализа литературных текстов: Моногр. М.: Эдиториал УРСС, 2012.

Орлов Ю. Н., *Осминин К. П.* Определение жанра и автора литературного произведения статистическими методами // Прикладная информатика. 2010. Т. 26, № 2. С. 95–108.

Свободина С. Ф. К вопросу о философской направленности и жанровых особенностях стихотворения А. С. Пушкина «Безверие» // Пушкинский музеум: Альманах. СПб., 2014. Вып. 6. С. 261–270.

Barakhnin V., *Kozhemyakina O.*, *Pastushkov I.* Automated determination of the type of genre and stylistic coloring of Russian texts // ITM Web of Conferences. 2017. Vol. 10. Art. 02001. 4 p.

Chawla N. V. Data Mining for Imbalanced Datasets: An Overview // Data Mining and Knowledge Discovery Handbook. Springer-Verlag, 2010. P. 875–886.

Cover T. M., *Hart P. E.* Nearest Neighbor Pattern Classification // IEEE Transactions on Information Theory. 1967. Vol. 13, iss. 1. P. 21–27.

Freund Y., *Schapire R. E.* A Short Introduction to Boosting // Journal of Japanese Society for Artificial Intelligence. 1999. Vol. 14, iss. 5. P. 771–780.

Friedman J. H. Stochastic gradient boosting // Computational Statistics and Data Analysis. 2002. Vol. 38, iss. 4. P. 367–378.

Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. Communications in Computer and Information Science. 2015. Vol. 542. P. 320–332.

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

Vapnik V. N. The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

Wolpert D. H. Stacked generalization // Neural Networks. 1992. Vol. 5, iss. 2. P 241–259.

Материал поступил в редколлегию 03.07.2017

Vladimir B. Barakhnin [1, 2], **Olga Yu. Kozhemyakina** [1]
Iliya S. Pastushkov [1], **Elena V. Rychkova** [1, 2]

[1] *Institute of Computational Technologies SB RAS*
6 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation

[2] *Novosibirsk State University*
1 Pirogov St., Novosibirsk, 630090, Russian Federation

bar@ict.nsc.ru, olgakozhemyakina@mail.ru, pas2shkov.ilya@gmail.com, helen@ict.nsc.ru

COMPUTER CLASSIFICATION OF RUSSIAN POETIC TEXTS BY GENRES AND STYLES

The automation of a complex analysis of poetic texts includes a task of identification of their generic and stylistic characteristics, which we consider as important attributes to determine the impact of low levels of verse (meter, rhythm, prosody, vocabulary, grammar) on the higher. To solve this particular task, the paper analyses the principles of formation of training sets which can be then used for refining automated algorithms proper to identify adequately the genres and styles of Russian poetic texts. To find out which algorithm of classification of poetic texts is the most accurate, we carried out a number of computational experiments with the corpus of lyceum poetry of A. S. Pushkin, including the method of assembling which involves the building of the compositions of algorithms, the advantage of this technique being that the errors of individual algorithms are mutually compensated. When assembling, we consider the algorithms in which the function referred to as algorithmic operator establishes a correspondence between a plurality of objects and a space of estimates, while another function, called the decision rule, establishes a correspondence between the space of estimates and the set of values of the objective function. As a result, the considered algorithms are given the form of a superposition of the algorithmic operator and the decision rule.

Our computational experiments based on single words, bigrams and trigrams from the poems of the Pushkin's training set allowed to test the most well-known methods of assembling, such as weighted voting, boosting and stacking. The algorithms developed for the task showed their efficiency as far as adequate identification of the genres and styles of Russian poetic texts is concerned. Also, it was found that even with simple classifiers based on lexical features or on n-grams it is possible to obtain good results. We established that on the basis of the criterion of maximizing of the minimum precision a multilayer perceptron should be used.

These computer algorithms can significantly simplify the work of experts investigating Russian poetic styles and genres.

Keywords: computer analysis of poetic texts, identification of genres and styles, classification algorithms.

References

Akhmanova O. S. Ocherki po obschey i russkoy leksigologii [Essays on General and Russian lexicology]. Moscow, Uchpedgiz, 1957. (In Russ.)

Barakhnin V. B., Kozhemyakina O. Yu. K probleme autentichnosti foneticheskogo analiza v svyazi s vozmozhnymi osobennostyami avtorskoy orfografii (na primere cheredovaniya okonchaniy -oy / -yy v lirike A. S. Pushkina) [The problem of authenticity of phonetic analysis in relation to the possible features of author's spelling (on the example of the alternation of the endings -oy / -yy in the poetry of A. S. Pushkin)]. *Vestnik Tomskogo gosudarstvennogo universiteta. Seriya: Filologiya* [*Bulletin of Tomsk State University. Philology*], 2016, vol. 13, № 2, p. 5–28. (In Russ.)

Barakhnin V. B., Kozhemyakina O. Yu. Ob avtomatizatsii kompleksnogo analiza russkogo poeticheskogo teksta [About the automation of the complex analysis of Russian poetic text]. *CEUR Workshop Proceedings*, 2012, vol. 934, p. 167–171. (In Russ.)

Barakhnin V., Kozhemyakina O., Pastushkov I. Automated determination of the type of genre and stylistic coloring of Russian texts. *ITM Web of Conferences*, 2017, volume 10, art.02001, 4 p.

Chawla N. V. Data Mining for Imbalanced Datasets: An Overview. *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag, 2010, p. 875–886.

Cover T. M., Hart P. E. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 1967, vol. 13, iss. 1, p. 21–27.

Freund Y., Schapire R. E. A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, 1999, vol. 14, iss. 5, p. 771–780.

Friedman J. H. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 2002, vol. 38, iss. 4, p. 367–378.

Grekhnev V. A. Lirika Pushkina. O poetike zhanrov [Lyrics of Pushkin. About the poetics of genres]. Gorkiy, Volgo-Vyatsky Book Publ., 1985. (In Russ.)

Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages. Analysis of Images, Social Networks and Texts. *Communications in Computer and Information Science*, 2015, vol. 542, p. 320–332.

Lestsova M. A. Opredelenie yadra i periferii zhanrov ody, pesni, poslaniya, elegii i epitafii na materiale angliiskih poetov-sentimentalistov XVIII veka [The determination of the core and the periphery of the genres of odes, songs, epistles, elegies and epitaphs on the works of English poets-sentimentalists of the XVIII century]. *Vestnik Chelyabinskogo gosudarstvennogo pedagogicheskogo universiteta* [*Bulletin of the Chelyabinsk State Pedagogical University*], 2014, iss. 4, p. 196–205. (In Russ.)

Lomonosov M. V. Predislovie o pol'ze knig tserkovnyh v rossiyskom yazyke [The preface about the advantages of Church books in Russian language]. *Lomonosov M. V. Complete Collection*. Moscow, Leningrad, Izd-vo AN SSSR, 1952, vol. 7, p. 585–592. (In Russ.)

Magomedova D. M. Filologicheskiy analiz liricheskogo stihotvoreniya [Linguistic analysis of a lyric poem]. Moscow, Publishing Center "Akademiya", 2004. (In Russ.)

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

Orlov Yu. N., Osminin K. P. Metody statisticheskogo analiza literaturnyh tekstov [Methods of statistical analysis of literary texts]. Moscow, Editorial URSS, 2012. (In Russ.)

Orlov Yu. N., Osminin K. P. Opredelenie zhanra i avtora literaturnogo proizvedeniya statisticheskimi metodami [The definition of the genre and the author of a literary work by statistical methods]. *Prikladnaya informatika* [*Applied Informatics*], 2010, vol. 26, no. 2, P. 95–108. (In Russ.)

Svobodina S. F. K voprosu o filosofskoy napravlennosti i zhanrovyh osobennostyah stikhotvoreniyaya A. S. Pushkina "Bezverie" [On the question about the philosophical orientation and genre peculiarities of poem of A. S. Pushkin "Unbelief"]. Pushkin Museum: An Almanac. St. Petersburg, 2014, iss. 6, p. 261–270. (In Russ.)

Vapnik V. N. The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

Wolpert D. H. Stacked generalization. *Neural Networks*, 1992, vol. 5, iss. 2, p 241–259.

For citation:

Barakhnin V. B., Kozhemyakina O. Yu., Pastushkov I. S., Rychkova E. V. Computer Classification of Russian Poetic Texts by Genres and Styles. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2017, vol. 15, no. 3, p. 13–23. (In Russ.)