УДК 003.349.081.3:004.932.75:004.65.021:004.85

А.А. Голубничий, А.Д. Яблонцева

База данных начертаний символов расширенных кириллических алфавитов современной тюркской письменности ExTL

Представлен алгоритм создания базы данных, содержащей начертания символов расширенных кириллических алфавитов современной тюркской письменности (ExTL). Рассмотрены методы и технологии искажения изображений при формировании набора данных. Итоговая база символов содержит 52920 изображений основной кириллицы (33 символа в двух регистрах), 29 символов расширенного кириллического алфавита в двух регистрах и 23 пунктуационных знака с использованием 14 типов искажений. Описан публичный интерактивный интерфейс к базе данных, созданный средствами языка программирования R и технологии Shiny.

Ключевые слова: *распознавание символов, тюркские языки, расширенный кириллический алфавит, язык R, наборы данных для машинного обучения*

DOI: 10.36535/0548-0027-2022-11-5

ВВЕДЕНИЕ

Для создания и тренировки моделей оптического распознавания символов часто используются базы данных, содержащие различные символы в рукописном формате, чаще всего это арабские цифры [1–6] или ограниченное количество символов латинского алфавита [7–9]. В некоторых исследованиях сделаны вполне успешные попытки создания аналогичных баз для иероглифов, имеющих также рукописную природу с последующей оцифровкой [10, 11]. При этом значительное количество изданной в середине двадцатого века литературы, отражающей самобытность тюркских народов, в настоящий момент находится исключительно в печатном виде, без электронных копий или же имеющая копии в виде изображений без дополнительного распознавания символов. Обработка такого рода сырых данных представляет большую сложность, поэтому создание базы данных для обучения моделей оптического распознавания печатных символов расширенных кириллических алфавитов имеет высокую актуальность.

ОБЪЕКТ И МЕТОДЫ ИССЛЕДОВАНИЯ

Для создания такой базы данных в качестве исходных были выбраны символы, составляющие письменность 23-х тюркских языков: азербайджанского, алтайского, башкирского, гагаузского, казахского, караимского, каракалпакского, карачаево-балкарского, киргизского, крымско-татарского, кумыкского, ногайского, сойотского, татарского, тофаларского, тувинского, туркменского, узбекского, уйгурского, хакасского, чувашского, шорского, якутского (долганского). Общий набор включенных символов составил 147, из них: 33 основные буквы кириллицы в двух регистрах, 29 букв расширенного кириллического алфавита в двух регистрах, 23 пунктуационных символа [12].

Для большего количества вариаций букв при построении моделей машинного обучения было принято решение использовать 6 шрифтов в четырех начертаниях без искажений и в 14 вариантах искажения изображений с последующим автоматизированным переводом и сохранением в файл сериализации формата RDS. Выбор этого формата, как базового для хранения набора данных, был обусловлен ранее полученными результатами обработки аналогичных данных на примере набора данных MNIST [13].

Набор шрифтов для построения системы был выбран исходя из следующих принципов: а) шрифт должен включать все символы основного кириллического алфавита и большую часть (не менее 90% символов) букв расширенного кириллического алфавита; б) при выборе следует использовать равное количество шрифтов с засечками и без; в) шрифты должны максимально различаться и обладать значительной распространенностью. Исходя их этих критериев для построения базы данных было выбрано 6 шрифтов: *Calibri, Times New Roman, Arial, Consolas, Cambria, Helvetica* и 4 типа начертаний: *regular, bold, italic, bold italic.*

Искажения получались, когда использовались 14 методов преобразования изображений, реализованных в свободном графическом редакторе *GIMP*. В таблице представлены описание и характеристики преобразований.

Наименование	Тип преобразования	Настройки
original	–	–
engrave	Гравировка	по умолчанию
mosaic	Мозаика	тип геометрии – квадрат, размер – 4
shift_1	Смещение	уровень – 1
shift_3	Смещение	уровень – 3
shift_5	Смещение	уровень – 5
wind	Ветер	по умолчанию
blast	Ветер	стиль – взрыв
blur_5	Размытие оптики	радиус – 5
blur_10	Размытие оптики	радиус – 10
pixel	Пикселизация	размер блока – 8
despeckle	Удаление пятен	радиус – 2
pick	Бросок	рандомизация – 10
spread	Рассеивание	размер – 10
glass	Стеклянные блоки	размер – 10

Рис. 1. Исходный набор символов

МЕТОДЫ И СРЕДСТВА ОБРАБОТКИ ИЗОБРАЖЕНИЙ. ПОСТРОЕНИЕ ИТОГОВОГО НАБОРА ДАННЫХ

Подготовка изображений происходила в несколько этапов: 1) создание собственно набора символов в нужном порядке; 2) выбор шрифта и начертания; 3) искажение согласно плану с сохранением результатов в виде отдельных файлов; 4) извлечение буквы с последующим центрированием. На рис. 1 приведен исходный набор символов в базовом начертании, используемый для дальнейших преобразований.

Большинство начертаний шрифтов, используемых для создания набора символов, имеют разную ширину букв (являются не моноширинными), и таким образом одной из сложностей при формировании набора символов стала проблема центрирования. Центрирование выполнялось автоматически, путем выбора крайних пикселей символов, которые отличались от белого цвета. Все изображения после центрирования были закодированы в квадраты размером 28 на 28 пикселей. Снижение размерности до данного уровня позволило использовать сравнительно небольшое количество входов для обучения алгоритмов, работающих с большими данными.

Набор данных, получивший название ExTL, был составлен из метки (*label*), указывающей на кодируемый символ, 784 переменных, отвечающих за кодирование каждого пикселя, за информацию о шрифте (*font*), начертании (*type*) и типе искажения (*distortion*). Общее количество вариаций символов в наборе данных составило 52920 значений, что соответствует 360 значениям на каждый из 147 символов. Сохранение служебной информации, такой как тип преобразования, необходимо для возможности исключения ряда искажений. В значительной степени это влияет на возможность распознавания символов, что дает определенную гибкость для настройки набора данных под конкретные нужды исследователя. Так, некоторый набор преобразований делает символы практически неотличимыми. На рис. 2 приведен пример одного из таких искажений.

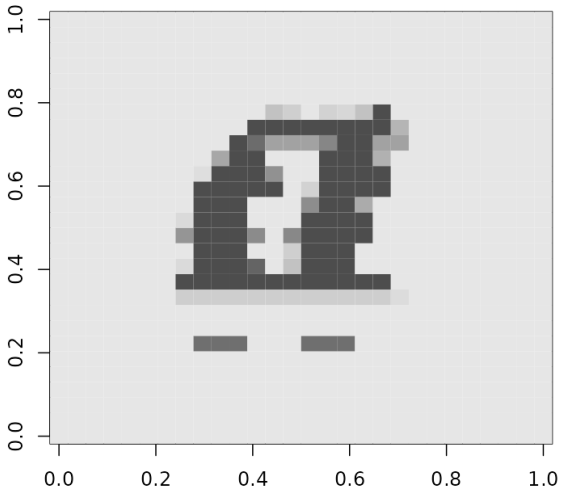

Рис. 2. Пример отображения буквы ***а*** с преобразованием типа «*engrave*» для шрифта *Times New Roman* в начертании полужирного курсива

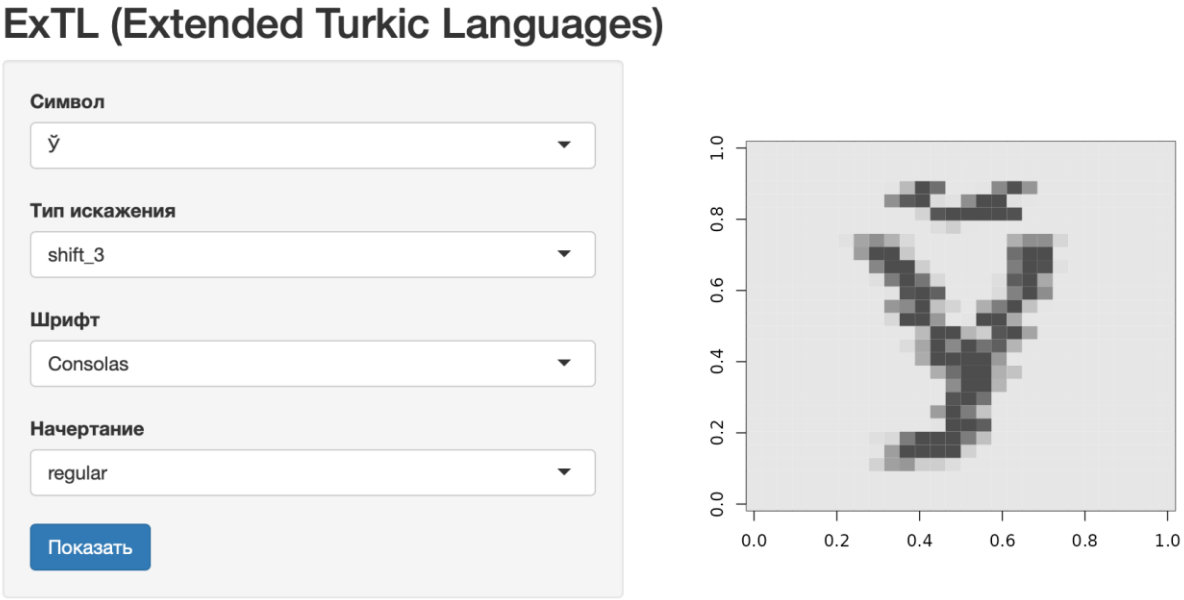

Рис. 3. Публичный веб-интерфейс для работы с базой данных ExTL [14]

РАЗРАБОТКА ИНТЕРФЕЙСА ДЛЯ ОБЗОРА НАБОРА ДАННЫХ

Для проверки корректности отображения символов мы представили публичный интерфейс, построенный с помощью языка программирования *R* и технологии *Shiny*, используемой для создания интерактивных приложений. Логика приложения предполагает выбор символа, его начертания и шрифта, а также применяемого искажения. После выбора всех настроек при нажатии кнопки «показать» строится рисунок, представляющий собой рендеринг исходной матрицы в градиенте серого. Интерфейс программного продукта, который можно использовать для предварительного обзора символов с целью дальнейшего выбора только части из них для построения итоговой модели обучения, приведен на рис. 3.

ПОСТРОЕНИЕ МОДЕЛИ РАСПОЗНАВАНИЯ СИМВОЛОВ С СОКРАЩЕННЫМ НАБОРОМ ДАННЫХ

Для оценки набора данных с точки зрения возможности применения алгоритмов машинного обучения была сформирована выборка, включающая вариации 10 символов (3600 строк исходного набора данных), с использованием букв как основной кириллицы (6 букв), так и расширенной (4 буквы). Размерность уменьшалась с целью применения таких алгоритмов, как «случайный лес», хорошо работающий на сравнительно небольшом количестве классов.

Исходный набор данных был разбит на тренировочную и тестовую выборки в соотношении 80% и 20%. Для построения модели использовался ансамбль из 50 деревьев.

```
Call:
 randomForest(x = train_set, y = train_labels, xtest = test_set,      ntree = 50)
               Type of random forest: classification
                     Number of trees: 50
No. of variables tried at each split: 28

        OOB estimate of  error rate: 2.81%
Confusion matrix:
    Ғ   Ң   Ө   Ү   Ъ   Ы   Ь   Э   Ю   Я class.error
Ғ 273   0   0   2   0   0   0   0   2   0  0.01444043
Ң   2 273   2   2   0   1   1   2   0   1  0.03873239
Ө   1   0 294   0   1   1   0   5   2   0  0.03289474
Ү   2   0   1 285   0   0   0   0   0   1  0.01384083
Ъ   0   0   0   2 282   0   7   0   1   0  0.03424658
Ы   0   0   2   1   1 277   4   0   1   0  0.03146853
Ь   0   0   0   0  10   2 270   0   0   0  0.04255319
Э   1   1   2   1   1   0   0 282   0   1  0.02422145
Ю   0   0   1   1   0   3   0   0 288   1  0.02040816
Я   1   1   0   1   0   0   0   0   5 275  0.02826855
```

Рис. 4. Результат работы алгоритма классификации с использованием технологии случайный лес на выборке исходного набора данных ExTL

Итоговый результат работы алгоритма приведен на рис. 4. Исходный код всей процедура анализа выглядел следующим образом:

```
extl_mini <- extl[extl$label %in% c
    ("Ъ","Ы","Ь","Э","Ю","Я","Ө","Ң","Ғ","Ү"),]
ind <- sample(1:3600, size = 720)
train_set <- extl_mini[-ind,]
test_set <- extl_mini[ind,]
train_labels <- as.factor(train_set[, 1]$label)
test_labels <- as.factor(test_set[, 1]$label)
head(train_labels, 20)
summary(train_labels)
library(randomForest)
rf <- randomForest(x = train_set, y = train_labels, xtest =
    test_set, ntree = 50)
rf
```

Оценивая общую эффективность работы алгоритма, с использованием промежуточных результатов, было получено значение 94,2% точности предсказания. Таким образом, ExTL в полной мере можно использовать в качестве тренировочного набора данных для отдельных алгоритмов машинного обучения.

СПИСОК ЛИТЕРАТУРЫ

1. Guyon I., Gunn S., Ben-Hur A., Dror G. Result analysis of the nips 2003 feature selection challenge // Advances in Neural Information Processing Systems. – MIT Press. – 2004. – Vol. 17. – URL: https://proceedings.neurips.cc/paper/2004/file/5e751896e527c862bf67251a474b3819-Paper.pdf.
2. Lecun Y., Bottou L., Bengio Y., Haffner P. Gradient-based learning applied to document recognition // Proceedings of the IEEE. – 1998. – Vol. 86. – № 11. – P. 2278-2324. – DOI: https://doi.org/10.1109/5.726791.
3. Kussul E., Baidyk T. Improved method of handwritten digit recognition tested on MNIST database // Image and Vision Computing. – 2004. – Vol. 22, Iss. 12. – P. 971-981. DOI: https://doi.org/10.1016/j.imavis.2004.03.008.
4. Alimoglu Fevzi, Alpaydin Ethem. Combining Multiple Representations for Pen-based Handwritten Digit Recognition // Turkish Journal of Electrical Engineering and Computer Sciences. – 2004. – Vol. 9. № 1. – URL: https://journals.tubitak.gov.tr/elektrik/vol9/iss1/1.
5. Tang E.K., Suganthan P.N., Yao X., Qin A.K. Linear dimensionality reduction using relevance weighted LDA // Pattern Recognition. – 2005. – Vol. 38, Iss. 4. – P. 485-493. DOI: https://doi.org/10.1016/j.patcog.2004.09.005.
6. Hong Y., Li Q., Jiang J., Tu Z. Learning a mixture of sparse distance metrics for classification and dimensionality reduction // Computer Vision (ICCV). – 2011 IEEE International Conference on. – IEEE. – 2011. – P. 906-913.
7. Botta M., Giordana A. and Saitta L. Learning fuzzy concept definitions // Second IEEE International Conference on Fuzzy Systems. – 1993. – P. 18-22. DOI: 10.1109/FUZZY.1993.327470.
8. Frey P.W., Slate D.J. Letter recognition using Holland-style adaptive classifiers // Machine learning. – 1991. – Vol. 6, № 2. – P.161-182. DOI: https://doi.org/10.1007/bf00114162.
9. Peltonen J., Klami A., Kaski S. Improved learning of Riemannian metrics for exploratory analysis // Neural Networks. – 2005. – Vol. 17, № 8-9. – P. 1087-1100. DOI: https://doi.org/10.1016/j.neunet.2004.06.008
10. Wang Da-Han, Liu Cheng-Lin, Yu Jin-Lun, Zhou Xiang-Dong. CASIA-OLHWDB1: A data-

base of online handwritten Chinese characters // 2009 10th International Conference on Document Analysis and Recognition. – 2009. – P. 1206-1210. DOI: https://doi.org/10.1109/ICDAR.2009.163.
11. Williams Ben H., Marc Toussaint, Amos J. Storkey. Extracting motion primitives from natural handwriting data // Lecture Notes in Computer Science. International Conference on Artificial Neural Networks. – 2006. – Vol. 4132. – P. 634-643. – Springer, Berlin, Heidelberg. DOI: https://doi.org/10.1007/11840930_66.
12. Голубничий А.А., Яблонцева А.Д. Формирование "белых списков" для создания универсальной системы распознавания тюркских языков, построенных на расширенных кирилличесих алфавитах // Научно-технический вестник Поволжья. – 2022. – № 4. – С. 67-70.
13. Голубничий А.А., Благосмыслова А.М. Оптимизация вычислений на уровне считывания и хранения данных (на примере MNIST) // Научно-технический вестник Поволжья. – 2022. – № 5. – С. 80-82.
14. ExTL (Extended Turkic Languages). – URL: https://extl.shinyapps.io/web_extl/.

Материал поступил в редакцию 21.09.22.

Сведения об авторах

ГОЛУБНИЧИЙ Артем Александрович – старший преподаватель Хакасского государственного университета им. Н.Ф. Катанова, г. Абакан
e-mail: artem@golubnichij.ru

ЯБЛОНЦЕВА Арина Дмитриевна – студент Хакасского государственного университета им. Н.Ф. Катанова, г. Абакан
e-mail: 91919919@mail.ru