2.3.1. (05.13.01)

А.А. Голубничий, А.Д. Яблонцева

Хакасский государственный университет имени Н.Ф. Катанова,
инженерно-технологический институт,
кафедра программного обеспечения вычислительной техники
и автоматизированных систем,
Абакан, artem@golubnichij.ru

ФОРМИРОВАНИЕ «БЕЛЫХ СПИСКОВ» ДЛЯ СОЗДАНИЯ УНИВЕРСАЛЬНОЙ СИСТЕМЫ РАСПОЗНАВАНИЯ ТЮРКСКИХ ЯЗЫКОВ, ПОСТРОЕННЫХ НА РАСШИРЕННЫХ КИРИЛЛИЧЕСКИХ АЛФАВИТАХ

В статье рассматриваются особенности построения универсальной системы оптического распознавания символов для тюркских языков, письменность которых построена на основе расширенных кириллических алфавитов. Проводится обзорный анализ основы для построения «белых списков», как набора символов, возможных для построения текстов исходных письменностей. Описываются два алгоритма построения «белых списков» с использованием полного и сокращенного набора символов. Уточняется создание языковых подмоделей для более точного распознавания символов с использованием пяти групп.

Ключевые слова: *тюркские языки, оптическое распознавание символов, Tesseract, мультиязычные системы распознавания символов, кириллические алфавиты.*

Введение

Официальным языком Российской Федерации является русский язык, при этом даже в преамбуле Конституции России утверждается важность многонационального народа России [1]. Одной из основ для сохранения культурного наследия этноса является язык и письменность, как инструмент ее сохранения. Все языки Российской Федерации относятся к 14 языковым семьям, некоторые семьи включают 2-3 языка, при этом тюркская языковая семья содержит 37 языков [2]. Сложность распознавания символов некоторых тюркских языков вызвана рядом факторов: наличие букв, имеющих схожие начертания, малое количество текстов для отдельных письменностей, плохое качество печати исходных текстов. При этом создание максимально точных систем оптического распознавания символов позволит в значительной степени ускорить и автоматизировать процесс оцифровки многих произведений и книг, играющих значительную роль в процессе сохранения культурного наследия и достояния народов России.

Структура тюркских языков и особенности тюркской письменности

Тюркская языковая семья содержит письменность, построенную на одном из четырех алфавитов с определенными расширениями. В настоящее время существует 23 письменности в качестве основы которых выступает кириллица: азербайджанская, алтайская, башкирская, гагаузская, казахская, караимская, каракалпакская, карачаево-балкарская, киргизская, крымскотатарская, кумыкская, ногайская, сойотская, татарская, тофаларская, тувинская, туркменская, узбекская, уйгурская, хакасская, чувашская, шорская, якутская (долганская). Помимо кириллицы письменность многих вышеобозначенных языков представлена на латинице (все кроме сойотской), арабице (14 языков) и иврите (караимская письменность).

Для построения универсальной системы распознавания символов особую роль играют исходные данные, так, например, при создании системы для распознавания двуязычных текстов особо важно включить необходимый дополнительный набор символов. В случае языков, построенных не на кириллице или латинице в чистом виде, важно включать символы, дополняющие исходные наборы. В качестве основного метода решения данной задачи выступают «белые списки», включающие всевозможные варианты для распознавания символов. Перед тем как формировать такие списки необходимо детально рассмотреть письменность всех ранее обозначенных языков (таблица 1).

Таблица 1 – Построение письменности отдельных языков с использованием кириллической базы

Письменность	**Буквы добавлены / исключены**	**Письменность**	**Буквы добавлены / исключены**
азербайджанская	Ғ Ә Ј Қ Ө Ү Һ Ҷ ' / Ё Й Ц Щ Ъ Ь Э Ю Я	сойотская	Ғ І Қ Һ Ң Ө Ү Ҷ Ә
алтайская	Ј Ҥ Ӧ Ӱ	татарская	Ә Ж Ң Ө Ү Һ
башкирская	Ғ Ҙ Ҡ Ң Ө Ҫ Һ Ә	тофаларская	Ғ і Ҕ Һ ҕ Ө Ҷ / Ц
гагаузская	Ӓ Ӝ Ӧ Ӱ	тувинская	Ң Ө Ү
казахская	Ә Ғ Қ Ң Ө Ұ Һ І	туркменская	Ж Ң Ө Ү Ә
караимская	Гъ Дж Къ Нъ Ӧ Ӱ Хъ / Ё	узбекская	Ў Қ Ғ Ҳ ' / Щ Ы
каракалпакская	Ә Ғ Қ Ң Ө Ў Ҳ	уйгурская	Ғ Һ Ж Қ Ң Ө Ү / Ё
карачаево-балкарская	Гъ Дж Къ Нг	хакасская	Ғ І Ң Ӧ Ӱ Ҷ
киргизская	Ң Ө Ү	чувашская	Ӑ Ӗ Ҫ Ӳ
крымско-татарская	Гъ Къ Нъ Дж	шорская	Ғ Қ Ң Ӧ Ӱ
кумыкская	Гъ Гь Къ Нг Оь Уь	якутская	Һ Ҕ Ө Ү
ногайская	Аь Нъ Оь Уь		

Формирование языковых моделей для распознавания символов

В целом, общая схема, используемая для дополнения стандартной кириллицы, при построении письменности тюркских языков представлена на рисунке 1.

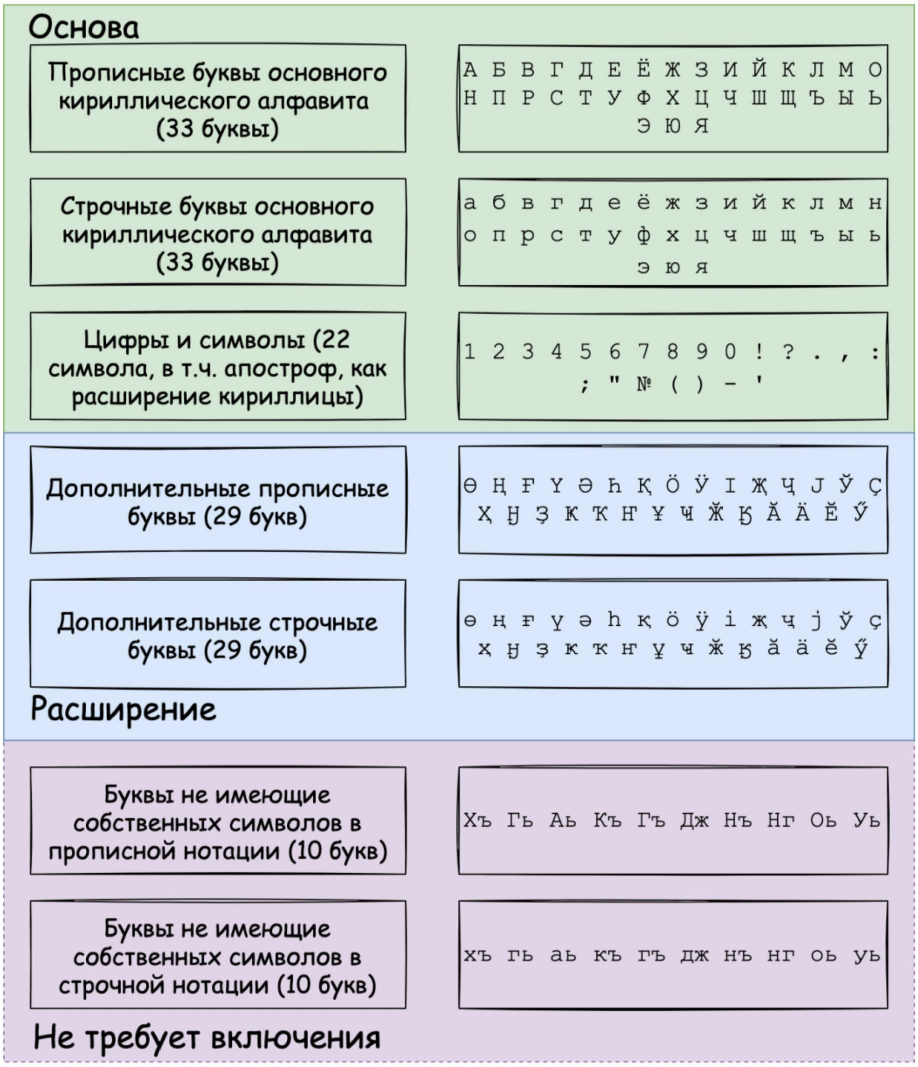

Рис. 1 – Структура символов, используемых при построении письменности языка

За исключением азербайджанской, практически все системы письменности строятся на полном наборе кириллических символов с добавлением некоторых дополнительных букв. Таким образом при обучении языковой модели, посредством технологии Tesseract, нет необходимости исключать отсутствующие буквы, а стоит сконцентрировать внимание на расширении списков возможных символов через формирование так называемых «белых списков». Алгоритм составления «белых списков» может быть построен по одному из двух вариантов: создание общего для всех письменностей списка, содержащего все буквы, используемые в алфавите и, создание ряда списков, использующих наиболее идентичные алфавиты для разных письменностей. Так, например, киргизская и тувинская письменность содержит полностью однотипные алфавиты, поэтому работы, посвященные оптическому распознаванию символов с помощью технологии Tesseract для тувинского языка [3], легко можно дополнить исследованиями для киргизского.

Построение белых списков осуществляется с использованием «основы» и «расширения», при этом буквы, не имеющие собственных символов и использующие двухбуквенное обозначение можно исключить из построения модели, как отдельные элементы. Данный подход возможен, по причине того, что при распознавании символов нейронные сети типа LSTM, применяемые в технологии Tesseract 5 помимо распознавания символов используют и контекст. Таким образом, непривычные сочетания для букв русской письменности, такие как аь, оь и др. в некоторой степени изменят общий контекст, но не будут требовать использование дополнительных символов при формировании «белых списков».

Второй подход для построения универсальных систем распознавания символов предполагает включение лишь некоторых символов для построения системы и использование более чем одного файла языковых моделей. Логичным в проведении анализа точности построения языковых моделей является разделение письменностей на следующие группы, приведенные в таблице 2.

Таблица 2 – Выделение групп для построения языковых моделей тюркских языков

Группа - Принцип формирования	**Письменности**
I. Без дополнительных букв. Письменности, в которых отсутствуют дополнительные буквы.	крымскотатарская, кумыкская, ногайская, карачаево-балкарская
II. Буквы с умлаутом, краткой и двойным акутом. Письменности в обозначении букв которых используются дополнительные элементы сверху (умлаут, кратка и др.).	алтайская, гагаузская, караимская, каракалпакская, хакасская, шорская, чувашская, узбекская
III. Буквы с нижними выносными элементами. Письменности в обозначении букв которых используются дополнительные элементы снизу (выносные элементы и крюк).	башкирская, каракалпакская, киргизская, сойотская, татарская, тофаларская, тувинская, туркменская, узбекская, уйгурская, чувашская, шорская, якутская
IV. Буквы с лигатурами. Письменности в обозначении букв которых используются лигатуры.	алтайская, башкирская
V. Перечеркнутые буквы и буквы со штрихом. Письменности в обозначении букв которых используются перечеркнутые буквы и буквы со штрихом.	азербайджанская, башкирская, казахская, каракалпакская, киргизская, сойотская, татарская, тофаларская, тувинская, туркменская, узбекская, уйгурская, шорская, якутская

Принцип разделения, как видно из таблицы, ориентируется на схожести новых добавляемых букв. Таким образом, обучаемая языковая модель будет построена на схожести элементов, что позволит сформировать более точные системы распознавания символов, за счет увеличения потенциального количества наборов для обучения.

Заключение

Предлагаемый принцип построения «белых списков» для создания систем распознавания символов позволит создать расширенную языковую модель для тюркских языков, построенных на основе расширенной кириллицы. Два предложенных метода для построения «белых списков» с полным перебором символов и с 5 малыми языковыми моделями, построенными на основе нескольких письменностей, потенциально могут достичь более точных результатов для построения систем типа OCR.

Список литературы

1. Конституция Российской Федерации (принята всенародным голосованием 12.12.1993 с изменениями, одобренными в ходе общероссийского голосования 01.07.2020) // Официальный интернет-портал правовой информации. URL: http://www.pravo.gov.ru, 04.04.2022.

2. *Дыбо А.В.* Хронология тюркских языков и лингвистические контакты ранних тюрков. — М.: Академия, 2004. — С. 766

3. *Голубничий А.А.* Разработка системы оптического распознавания символов тувинского алфавита / А. А. Голубничий, А. Д. Яблонцева, В. А. Мясоедова // Научно-технический вестник Поволжья. – 2021. – № 12. – С. 156-158.