# ****Введение****

На сегодняшний день распознавание рукописного текста является весьма актуальной задачей. Решение данной проблемы позволило бы автоматизировать бизнес процессы множества компаний. Одним из наглядных примеров могут служить почтовые компании, где остро стоит вопрос сортировки большого объема писем и посылок. Целью проекта, в рамках которого пишется данная статья, является распознание рукописных адресов письменной корреспонденции.

Задачу распознавания адреса можно считать упрощенной задачей распознавания текста, потому что адреса используют ограниченный набор слов, и этот набор существенно меньше, чем множество всех слов, принадлежащих тому или иному языку. Это в разы уменьшает количество необходимых данных и позволяет сделать некоторые допущения. Данная статья посвящена сбору данных, которые использувются для обучения модели.

В первой главе мы более подробно опишем отличие общей задачи распознавания текста от нашей. Во второй главе приводятся примеры существующих баз данных рукописных текстов. В третьей главе детально разбирается процесс сбора данных. Далее, рассмотренны вопросы хранения данных. В заключении, приведены выводы и дальнейшая работа по данной задаче.

## ****1. Отличие распознавания адреса от общей задачи распознавания текста****

Любой язык содержит в себе огромное количество слов. К примеру, словари русского и казахского языков в среднем регистрируют более 100 000 слов, а оксвордский словарь английского языка – более 300 000 слов. В связи с этим собрать исчерпывающую базу данных рукописных слов, которая включала бы в себя все слова с большой вариацией почерка, представляется практически невозможным. Другими словами, всегда найдется слово, которое система не сможет распознать.

В отличие от общего случая с текстом, адрес может включать в себя слова из более ограниченного множества специфичных для адреса слов. В большинстве случаев, адрес - это короткий блок из 5-15 слов, в котором указываются страна, штат/провинция/город, улица, номер дома, почтовый индекс, фамилия и имя отправителя/получателя. Причем, фамилия и имя отправителя/получателя в решении задачи сортировки корреспонденции никаким образом не участвуют, т.к. не несут никакой информации о направлении письма, поэтому они могут быть опущены. Кроме того, порядок слов в адресе не имеет значения, так как любая перестановка элементов адреса не изменит конечного пункта назначения.

Дополнительно, в адресе наблюдается некая иерархия. К примеру, если в адресе в качестве страны указан Казахстан, то закономерно предположить, что далее последует один из населенных пунктов на территории РК. Таким образом пространство поиска может быть значительно сокращено. Подобная иерархическая зависимость позволяет рассматривать более гибкие и легко масштабируемые архитектуры системы. Например, мы можем создать систему, которую «научим» распознавать название страны назначения. Официально в мире насчитывается 197 стран. Далее, для каждой станы мы обучим модели распознавания названия городов. Итак, по мере необходимости система будет расширяться, причем количество необходимых моделей не будет превышать 197.

## ****2 Существующие базы данных рукописного текста****

Существует множество датасетов с рукописным текстом. Большинство из них содержат текст на английском языке. Эти датасеты не ограничены контекстом, то есть генеральная популяция контекстов апроксимируется случайной выборкой. Одними из ярких примеров датасетов являются:

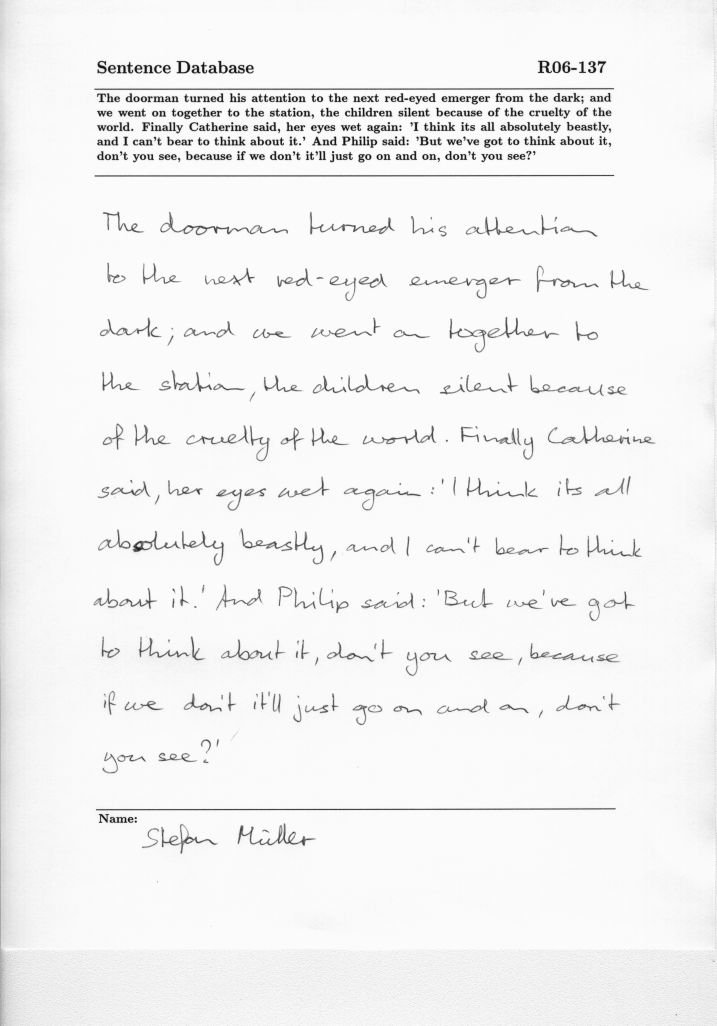
1. IAM Handwriting Database
2. CEDAR
3. **CVL-Database**
4. **ICDAR2013-2017**
5. NIST
6. CENPARMI
7. UNIPEN
8. IRESTE On/Off (IRONOFF) Dual Handwriting Database

Аналогов на русском и казахских языках не существует. Мы адаптировали общие принципы сбора и хранения данных, описанные в [Zimmermann] и использованные для сбора данных IAM Database. Описание данной базы данных смотрите ниже.

**2.1. Описание IAM Database**

**Создатели данной базы данных использовали 13 категорий текста. В каждом тексте около 2000 слов, поделенных на 3-6 фрагментов по 50 слов. Каждая форма содержит 4 области, разделенные горизонтальными линиями:**

1. **Заголовок и идентификационный номер (категория-индекс текста-индекс предложения)**
2. **Печатный текст**
3. **Пустая зона для рукописного текста**
4. **Зона для имени (необязательно)**

****

**Рис 1. Образец формы из базы данных IAM.**

**Разделяющие линии используются для сегментации зон и определения пространственных деформаций листа при сканировании. После сегментации текст режится по линиям. У каждого слова в печатном тексте есть своя метка. Эти метки копируются для всего фрагмента, и словам в рукописном варианте, которые находятся в строке, в ручную присваиваются метки.**

**Результатом размечивания (labeling) является ASCII файл, который содержит ярлыки для каждой строчки печатного и рукописного текста. В файл не включается пространственная информация. Каждый участник опроса заполнял от одной до трех форм разных категорий.**

По данным с сайта известно, что в опросе учавствовали 657 человек. В общей сложности по шаблону, изображенному на Рис. 1, было собрано 1539 страниц текста. Эти страницы сегментированы на 5685 предложений, 13353 аннотированных строчек и 115 320 слов.

# ****2.2. Описание CVL-Database****

**Количество людей, заполнивших формы, равно 310. База данных использовала 7 различных текстов (1 на немецком и 6 на английском). Документы сканировались с разрешением 2499x3520 пикселей. Каждой форме присвоены идентификационный номер заполнившего форму человека, а также номер текста.**

**Более подробная информация о том, как авторы собирали данные и сегментировали текст, отсутствует.**

# ****2.3. Описание ICDAR2013 dataset****

ICDAR2013 использовалась в рамках конференции 2013 года. Она содержит 150 сканов рукописного текста на греческом и английском языках. Кроме того, 50 изображений содержат текст, написанный на бенгальском.

# ****3 Сбор и хранение данных****

Этап сбора данных является одним из наиболее трудоемких и затратных этапов. Наша основная задача – это максимально упростить и автоматизировать его. В контексте распознавания рукописного адреса необходимо определить множество ключевых слов, которые могут встречаються в адресе, и собрать следующее:

* образцы рукописного алфавита на кириллице и латинице

## **рукописные образцы ключевых слов** на казахском, русском, английском языках

## **сканы лицевой стороны конвертов**

Далее необходимо будет сделать аннотацию собранных данных, т.е. разметить, и синтезировать новые образцы уже из существующих, посредством применения различных геометрических и фотометрических трансформаций (data augmentation).

Анализ сканов конвертов показал следующее:

1. Расположение текста на конверте произвольное
2. Формы и размеров конвертов разные
3. **На конвертах могут присутствовать рисунки**
4. Текст рукописный
5. Ошибки в тексте (с/без исправлений)
6. Цвет не фиксированный
7. **Могут быть печати с адресом**
8. **Печати могут быть разных форм (круглые, прямоугольные) и размеров**
9. **Печати могут быть плохо видны (размазаны, печать «сухая») или хорошо видны.**
10. **Может присутствовать паразитный текст и печати (т.е. текст, отличный от отправителя и получателя)**
11. **На конвертах могут присутствовать марки**
12. **На марке может быть тоже текст.**
13. **Ориентация текста любая (например, вес письма, время отправления сбоку поперек).**
14. **Печати могут быть поставленны поверх текста**
15. **Фотографии/сканы могут быть сделаны с перспективой (неортогональное положение относительно камеры).**
16. **Текст может быть сокращен**
17. **Расстояние между словами и между строчками может быть маленьким или большим**
18. **Может быть перенос слов с одной строки на другую**
19. **Строчки текста могут быть неровными (с наклоном, с выпадением отдельных букв, слов, вставками слов)**

## ****3.1. База данных ключевых слов****

Для начала, мы рассматриваем входящую в РК корреспонденцию, поэтому список ключевых слов содержит следующие названия:

* Областей
* Городов
* Сел
* Поселков
* Районов
* Улиц

Дополнительная информация такая, как:

* Индексы
* Телефоны
* Фамилии
* Названия компаний

в базу данных не входит.

## ****3.2.**** Образцы рукописного алфавита

Существует два принципиальных подхода распознавания текста: по символьное распознавание (Optical Character Recognition, OCR) и распознавание по словам (Optical Word Recognition, OWR).

При OCR для обучения модели требуется датасет, содержащий рукописные образцы всех символов алфавита для каждого языка. Здесь важно для каждого языка составить отдельные формы, так как набор букв различных алфавитов может сильно различаться.

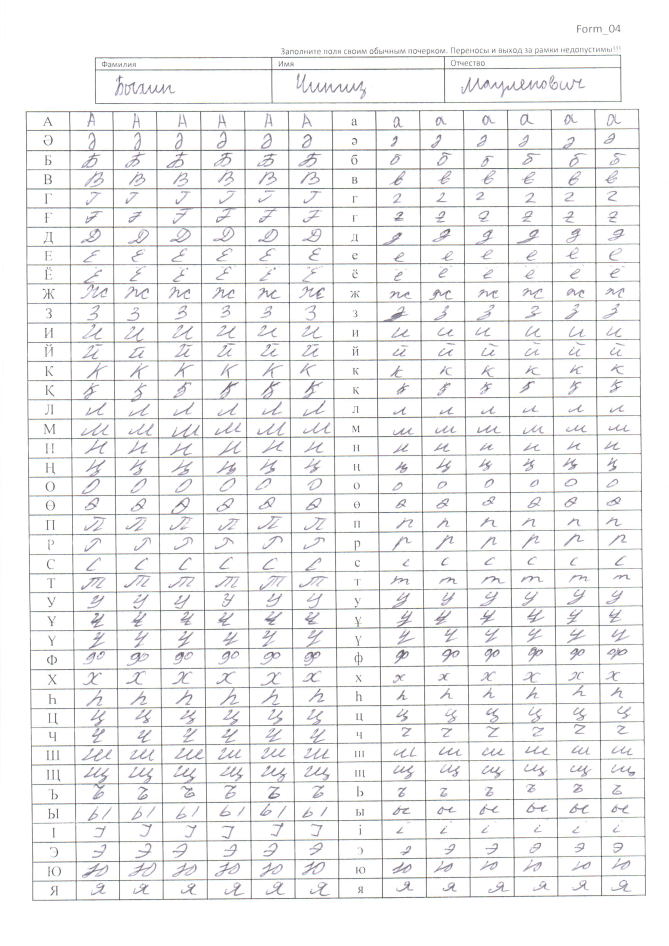


Рис 2. Один из первых вариантов формы для сбора рукописных образцов алфавита кириллицы.

Далее, для последующего обучения и тестирования модели понадобятся рукописные образцы целевых слов. Пример одной из форм для сбора образцов слов приведен на Рис. 3.

****

Рис 3. Пример формы для сбора данных

Данная форма спроектирована таким образом, чтобы была возможность легко идентифицировать и сегментировать по словам. Для идентификации формы в правом верхнем углу располагается маркер. Для упрощенния процесса сегментации вся форма разделена горизонтальными и вертикальными линиями, которые позволяют довольно легко восстановить структуру документа, и, соответственно, пространственное положение слова. Слова индексируются (аннотируются) согласно своего положения в таблице.

Для того чтобы вырезать слова из формы производятся следующие действия (preprocessing):

1. фильтрация форм с целью усиления границ таблицы
2. определение контуров таблицы
3. определение и компенсация угла поворота
4. исключение артефактных линий
5. сортировка форм по id (маркеру)
6. потоковое разделение форм на слова
7. наименование и хранение слов

После того, как области изображения, соответствующие ячейкам со словами сегментированы, они могут содержать некоторые краевые артефакты. Например, к артефактам можно отнести линии таблицы, вырезанные вместе с ячейкой, или части слова из соседней ячейки (Рис. 4). Мы устраняем эти артифакты, строя вертикальную и горизонтальные гистограммы и отсекая части отделенно локализированные ближе к краям ячейки.

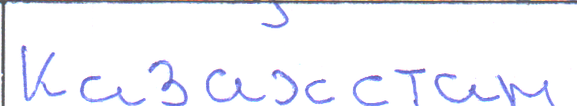


Рис 4. Пример вырезанной из формы области со словом. По краям видны ярко выраженная линия ячейки и кусочек буквы из соседней области.

Однако, не всегда возможно устранить все артефакты. Ниже приведены некоторые аспекты, затрудняющие дальнейшую обработку сегментированного слова:

1. буквы могут быть не соединены между собой
2. буквы могут персекаться с артефактами
3. положение букв и их крупность значительно варьирируются от слова к слову
4. буквы могут быть написаны разным цветом (синим, черным, красным)

В связи с этим, мы разработали систему рекоммендаций, которая позволяет упростить процесс выделения областей со словами из формы.

Во-первых, мы предлагаем заполнять форму с использованием синей ручки**.** Это позволит системе отличать слово от границ таблицы на уровне цвета. Например, переведя изображение из RGB в HSV, мы получаем инвариантное по отношению к освещению цветовое представление объектов. В данном цветовом пространстве синие остается синим вне зависимости от яркости и интенсивности изображения.

Во-вторых, устранение частей слов из соседних ячеек порой невозможно без искажения целевого содержимого данной ячейки, поэтому при заполнении формы желательно, чтобы испытуемый не выходил за границы ячейки.

Далее можно произвести нормализацию слов, то есть:

* 1. привести все изображения слов к одному размеру
  2. центрировать слово на изображении
  3. уменьшить размер изображения

Однако, современные модели, такие как для конволюционных и рекуррентных нейронных сетей данный этап может быть опциальным, так как их тренировка предполагает данные, содержащие любые геометрические трансформации объектов, изображенных на картинках.

**3.3. Сбор сканов лицевых сторон конвертов**

Любая задача машинного обучения с учителем требует входных размеченных данных, на которых можно было бы обучить модель. Причем, нам необходимо обучить как минимум две модели: одну для определения областей снимка, где располагается текст, а другую для распознавания слов.

Следовательно, мы должны собрать как минимум два набора данных: снимки реальных конвертов (Рис. 5), подобные тем, с которыми нам предстоит работать, а другой набор – множество рукописных целевых слов (описан в предыдущей главе).

Мы предполагаем, что первый набор данных (снимки конвертов) мы получим от КазПочты. В данный момент ведутся интенсивные переговоры по поводу этого вопроса.



Рис 5. Снимки конвертов под разным углом.

Каждый конверт содержит свой адрес

## ****3.4. Способы сбора данных****

**Человек, который согласился предоставить образец своего почерка, должен будет сделать следующее:**

1. **Получить целевой текст.**
2. **Написать его на чистом листе A4.**
3. **Сфтографировать.**
4. **Выслать нам.**

**Это можно организовать следующим образом:**

* **email**
* **разработка и запуск telegram-бота**
* **установка оборудования для сканирования адресов в пунктах приема писем**
* **Сканирование старой имеющейся корреспонденции**

**4. Заключение**

Сбор рукописных образцов ключевых слов и снимков конвертов будет продолжаться. Параллельно идет аннотация конвертов и проверка различных метрик для оценки погрешности распознования. Для того, чтобы следующие артефакты не были помехой нам нужно собрать как можно больше размеченных данных.

**5. Референсы**

1. <http://www.cedar.buffalo.edu/Databases/>
2. <http://www.fki.inf.unibe.ch/databases/iam-handwriting-database>
3. <https://cvl.tuwien.ac.at/research/cvl-databases/an-off-line-database-for-writer-retrieval-writer-identification-and-word-spotting/>
4. <http://users.iit.demokritos.gr/~nstam/ICDAR2013HandSegmCont/>
5. <http://www.irccyn.ec-nantes.fr/~viardgau/IRONOFF/ICDAR99.htm>
6. [Zimmermann] M. Zimmermann, H. Bunke, Automatic Segmentation of the IAM Off-line Database forHandwritten English Text, ICPR 2002