**Уровень абстрации 1**

Основная задача проекта – это сортировка входящей в Казахстан письменной корреспонденции.

**Уровень абстрации 2**

**Уровень абстрации 3**

Конверты очень сильно отличаются друг от друга: по размеру, цвету и содержанию. К примеру, на конвертах может быть разное количество марок, паразитный текст или положение областей получателя-отправителя могут быть смещены или вообще отсутствовать. Следовательно, невозможно обеспечить 100% точность работы системы. В связи с этим, в случае некорректного автоматического срабатывания системы, мы предоставляем оператору возможность вручную указывать область снимка, где находится область получателя.

**Уровень абстрации 4**

Письменная корреспонденция может быть как внутренней, так и международной, поэтому нам необходимо рассматривать несколько языков: казахский, русский, английский.

Для большей точности рекоммендуется иметь отдельную модель распознавания для каждого языка.

Допустим мы разбили область снимка с данными получателя на слова. Предположим, что в этом наборе слов обязательно есть название страны. В случае, если страна не указана, то по умолчанию корреспонденция считается внутренней.

Внутренняя переписка может вестись на казахском и русских языках. Переписка по странам СНГ может вестись на русском языке. Международная переписка – на английском.

С учетом этого допущения, для определения языка нам необходимо иметь словарь ключевых слов для каждого языка. В данном случае, ключевыми словами будут считаться названия стран ( слов если учитывать сокращения и т.д.).

Следовательно, мы пробегаемся по всем картинкам, распознавая слова с использованием одной из моделей распознавания, и сравниваем с каждым словом в базе данных ключевых слов. С использованием определенной метрики выбираем наиболее вероятную модель распознавания языка.

Допуская, что модель распознавания корректно выбрана и страна корректно определена. Для каждой страны мы храним словарь ключевых слов, который включает в себя названия провинций, областей, городов (<1000 слов). Этого должно быть достаточно для сортировки писем для межгородских сообщений.

Следовательно, для коррекции результатов распознавания мы можем находить степень соответсвия ключевым словам и заменять соответствующими словами.

**2020 год**

этот год в основном работали над моделями, результаты сравнения описали в приложенных статьях. По Интерфейсу ничего нового нет, только то что реализовал в прошлом году Жулдызжан. Про интерфейс можешь найти в отчете прошлого года.

Запустили новый метод сбора данных с использованием телеграм бота. Отсканировали экзаменационные листы студентов, сейчас около 2000 страниц А4, планируется еще около 4000. Сегментировали слова и и запустили аннотацию. До этого делали телеграм бот, где был обратный принцип, мы отправляли слова пользователям, а они должны были прикреплять фото с рукописным текстом. Но этот метод не пошел, так как очень много телодвижении.

2000 листов сегментировали 180000 слов