Пост-обработка результатов OCR распознавания, использующая частично определенный синтаксис

Д. Л. Шоломов, В. В. Постников, А. А. Марченко, А. В. Усков

В работе рассматривается задача пост-обработки текстовых полей ввода на формах и структурированных документах. Часто формы содержат поля с довольно свободным синтаксисом. Тем не менее, на документах с которыми мы имеем дело, поля со свободным синтаксисом практически всегда содержат характерные синтаксические конструкции. В этом случае нашей задачей является интерпретация такого рода конструкций с целью улучшения качества распознавания. В статье описан так называемый PDS подход использующий Частично Определенный Синтаксис, который в частности был успешно применен при распознавании полей ввода на платежных документах Центрального Банка РФ, а также при вводе анкет Пенсионного Фонда РФ. В статье приводится процедура Автоматического Выделения Синтаксиса (ASE) и MCHSR алгоритм быстрого нечеткого отображения текстовой строки применяемый при отображении словарей. Также приводятся результаты пост-обработки поля «Назначение платежа» на Платежных Поручениях ЦБРФ и результаты сравнения PDS подхода с Синтаксическим подходом на примере распознавания поля «Почтовый адрес» на анкетах Пенсионного Фонда РФ.

1. Введение

Системы распознавания и идентификации структурированных документов широко используются в сфере информационных технологий. Современные системы оптического распознавания достигли такого качества, которое вряд ли может существенно быть улучшено без использования информации о контексте распознавания. Часто поля ввода на формах имеют некоторую синтаксическую структуру либо серьезные семантические связи с другими полями. Информацию такого рода необходимо использовать в процессе распознавания для получения качественно более высоких результатов распознавания. Известные подходы контекстной

пост-обработки [2–4, 6] включают статистические и лингвистические методы использующие Скрытые Марковские Модели (СММ) [9], конечные автоматы, нейронные сети, N-граммы символов и слов, алгоритмы нечеткого отображения строк [2]. Также используются методы использующие специальную внешнюю информацию, комбинированные методы, а также подходы основанные на эвристиках. Часто известные методы не подходят для обработки полей структура которых не сводится к словарю или естественному языку. Структура последних либо слишком жесткая, либо слишком свободна.

В нашей предыдущей работе был рассмотрен Синтаксический Подход (SA) [5] в котором синтаксис поля определялся синтаксическими диаграммами, и использовалась специальная ОП-процедура для нахождения оптимального пути проходящего одновременно через синтаксическую диаграмму и сеть результатов распознавания (АР-цепь). На примере распознавания рукопечатного поля «Почтовый адрес» было показано, что SA существенно улучшает результаты распознавания. К сожалению, процесс точного синтаксического описания поля часто является довольно сложным и трудоемким процессом. Поэтому требуется иметь более легковесный подход к пост-обработке таких полей, которые тяжело синтаксически описать полностью, либо задать разумный синтаксис практически невозможно. В таком случае оправдано использование частично-определенного синтаксиса, который описывает только часто встречающиеся текстовые структуры. В работе предлагается подход, использующий Частично Определенный Синтаксис (PDS подход), который успешно был применен для распознавании целого ряда полей на платежных документах Центрального Банка РФ и анкетах Пенсионного Фонда РФ. Подход использует облегченный последовательно-параллельный синтаксис и автоматическую процедуру его построения (ASE процедуру). Благодаря ASE процедуре, разработка специальных функций пост-обработки занимает гораздо меньшее время. PDS подход и ASE процедура рассматриваются более подробно в главе 4.

2. Задача пост-обработки

Обычно процесс распознавания формы включает стадии предобработки изображения формы, привязки шаблона формы, локализации полей ввода, сегментации строки и распознавания символов [1]. После того как поле распознано, предварительные результаты распознавания сохраняются в сеть альтернатив распознавания АР-сеть (см. рис. 1).

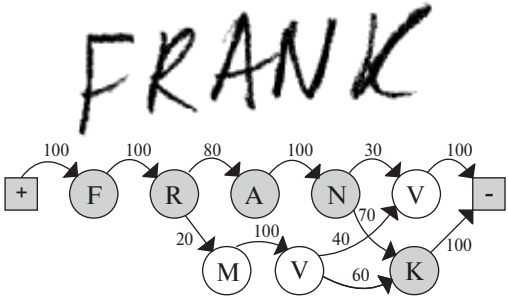


Рис 1. Пример АР-сети

Вершины АР-сети соответствуют альтернативе (варианту) распознавания символа, а ребра — оценке перехода от одной альтернативы к другой. Начальная и конечная вершина обозначаются знаками плюс и минус соответственно. АР-цепь содержит как информацию о распознанных символах, так и о вариантах сегментации строки. Каждому пути через АР-сеть соответствует текстовое значение, собранное из символов в вершинах сети. Задача состоит в нахождении оптимального пути в АР-сети с учетом синтаксиса поля. Часто необходимо нормализовать распознанное значение. Практически всегда необходимо оценить степень достоверности распознанного значения и локализовать его сомнительные фрагменты.

Процедура поиска оптимального пути в АР-сети — довольно трудная оптимизационная задача т. к. АР-сеть может состоять из сотен вершин, а синтаксис поля может быть весьма сложным.

В случае печатного либо рукопечатного заполнения поле в большинстве случаев хорошо сегментируется. Поэтому удобно представлять результаты распознавания в виде АР-матрицы (см. Рис. 2). Ячейки матрицы содержат альтернативу распознавания, а также ее оценку. Столбцы матрицы соответствуют знакоместам.

3. PDS подход

Подход, использующий частично-определенный синтаксис допускает пост-обработку полей, которые трудно либо невозможно описать Синтаксическими Диаграммами (СД) полностью. Но в то же время такие поля могут содержать типичные для поля стабильные синтаксические конструкции. Идея состоит в том, чтобы задать синтаксис только для типичных конструкций, при этом полностью описывать синтаксис поля не требуется.

j_{80}	4_{80}	n_{50}	e_{70}
i_{20}	u_{20}	u_{30}	c_{30}
		4_{20}	

0 1 2 3 4

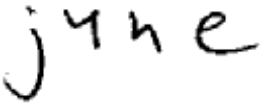
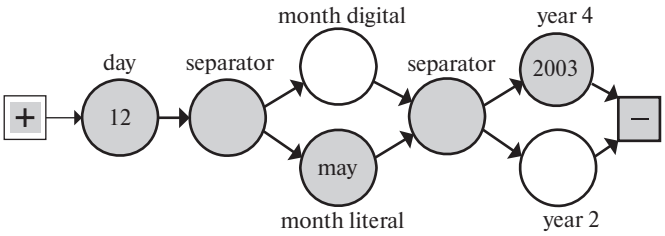


Рис 2. Пример AP-матрицы

Обычно контекстно-свободный синтаксис задается грамматиками, например в форме Бэкуса—Наура (БНФ), либо при помощи синтаксических диаграмм (СД). Мы используем параллельно-последовательный частично-определенный синтаксис, задаваемый посредством параллельно-последовательных СД (см. рис. 3а). Мы используем последовательно-параллельный синтаксис т. к. при этом проще автоматически строить синтаксис на основании обучающих данных. Последовательно-параллельные СД описывают широкий спектр полей, при этом существует хорошая синтаксическая нотация аналогичная нотации арифметических выражений, но только с операциями ИЛИ и КОНКАТ (или и конкатенация), определяемых производной БНФ грамматикой (см. ниже).



a)

<+><#1−2> <month>|<#2> <#2,4><−> *b)*

Рис 3. (a) Синтаксическая диаграмма даты,
(b) Синтаксическая нотация даты

Синтаксическая диаграмма PDS представляет собой направленный ациклический граф, содержащий начальную и конечную вершины (см. рис. 3а)

Вершины СД могут выть как терминальными, так и составными. Каждая составная вершина также определяется при помощи СД. Будем рассматривать терминальные вершины как вершины одного из следующих типов: статический текст, число, словарное слово, слово, последовательность символов из определенного алфавита, начальная и конечная вершина.

Зададим частично-определенный синтаксис (PDS) посредством следующей производной БНФ грамматики:

```
1:  <DIGIT>         :=   "0" | … | "9"
2:  <LETTER>        :=   "a" | … | "z"
3:  <PUNCT>         :=   "." | "," | " " | ":" | ";" |…
4:  <CHARACTER>     :=   <DIGIT> | <LETTER> | <PUNCT>
5:  <NUMBER>        :=   <DIGIT>+
6:  <WORD>          :=   <LETTER>+
7:  <STRING>        :=   <CHARACTER>+
8:  <RANGE>         :=   <NUMBER> "-" <NUMBER>
9:  <ENUM>          :=   <NUMBER> ("," <NUMBER>)*
10: <LEN_RESTRICT>  :=   <RANGE> | <ENUM>
11: <LEXICON_ID>    :=   <STRING>
12: <CTEXT_NODE>    :=   <STRING> | "'" <STRING> "'"
13: <LEX_NODE>      :=   "<" <LEXICON_ID> ">"
14: <NUM_NODE>      :=   <#> | "<#" <LEN_RESTRICT> ">"
15: <WORD_NODE>     :=   <$> | "<$" <LEN_RESTRICT> ">"
16: <CHRSEQ_NODE>   :=   "<*" <STRING> ">" |
                         "<*{" <STRING> "}"
                         <LEN_RESTRICT> ">"
17: <START_NODE>    :=   "<+>"
18: <FINISH_NODE>   :=   "<->"
19: <TERM>          :=   <CTEXT_NODE> | <LEX_NODE> |
                         <NUM_NODE> | <WORD_NODE> |
                         <CHRSEQ_NODE> | <START_NODE> |
                         <FINISH_NODE>
20: <OR_EXPR>       :=   <TERM> | <BRK_EXPR> |
                         <OR_EXPR> "|" <OR_EXPR>
21: <BRK_EXPR>      :=   "(" <CAT_EXPR> ")" |
                         "(" <OR_EXPR> ")" |
                         "(" <BRK_EXPR> ")"
22: <CAT_EXPR>      :=   <OR_EXPR> " " <OR_EXPR>
23: <PDS_EXPR>      :=   <CAT_EXPR> | <OR_EXPR> | <BRK_EXPR>
```

Синтаксические правила 1–18 определяют следующие основные типы терминальных вершин:

- **<CTEXT_NODE>** (статический текст) — задается текстовым значением, например **текст** либо **'текст из нескольких слов'**.
- **<LEX_NODE>** (слово из словаря) — задается идентификатором словаря в угловых скобках, например **<street>**.
- **<NUM_NODE>** (число) — задается знаком **<#>** либо **<#length>,** например, **<#2–4>** либо **<#1,3,8>**.
- **<WORD_NODE>** (слово) — знаком **<$>** либо **<$length>**.
- **<CHRSEQ_NODE>** (последовательность символов) — задается как **<*alphabet>** либо **<*{alphabet}length>**, например. **<*1234abc>** либо **<*{1234abc}2–8>**.
- начальная и конечная вершины как **<+>** и **<->** соответственно.

Правила (19–23) определяют синтаксис PDS выражений, которые могут содержать операции ИЛИ, КОНКАТ и скобки. Операция ИЛИ имеет более высокий приоритет, чем конкатенация. Пример PDS выражения задающего дату приведен на рис. 3b.

Пусть каждая устойчивая текстовая конструкция задается PDS выражением. Полученный список PDS выражений полностью описывает синтаксис часто встречаемых текстовых фрагментов. В этом случае будем говорить, что частично-определенный синтаксис (PDS) задан. PDS процедура — это метод нахождения оптимального пути в AP-цепи, также удовлетворяющего частично-определенному синтаксису. В случае если фрагмент AP-цепи соответствует некоторому PDS выражению, он обрабатывается соответствующим образом. А фрагменты, которые не соответствуют PDS выражениям, обрабатываются общим алгоритмом постобработки, например, при помощи n-грамм либо коррекцией по словарю. Также можно взять в качестве финального текстового значения наилучшие альтернативы символов по знакоместам.

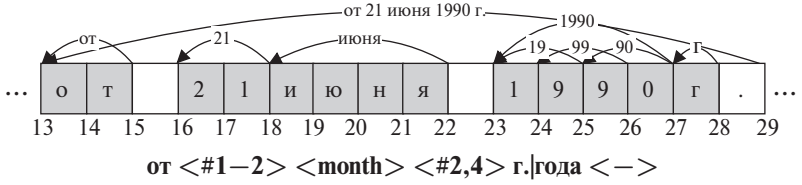


Рис 4. (a) Ребра отображенные на AP-цепь,
(b) PDS выражение отображенное на интервал [13–29]

PDS процедура состоит из следующих стадий:

1. Отображение статического текста и словарных вершин

Для единообразия список статических текстов будем рассматривать как словарь ключевых слов. Вначале все словари инициализируются. После этого слова из каждого словаря отображаются на AP-цепь. Словари лексикографически упорядочены, что позволяет использовать MCHSR — быструю процедуру отображения словаря на AP-цепь основанную на динамическом программировании. Слово из словаря отображается на AP-цепь с некоторой оценкой. Если оценка превосходит определенный порог, то так называемое ребро (см. рис. 4а) сохраняется в специальном контейнере ребер. Ребро содержит начальную и конечную позиции AP-цепи определяющие отрезок, на который отобразилось слово, тип ребра (в данном случае — словарное ребро), отображенное слово, оценку отображения и идентификатор словаря. На рис. 4а ключевое слово «от» отобразилось на интервал [13–15] а слово «июня» из словаря month — на интервал [18–22]. Оценка отображения $s(w)$ слова *w* определяется следующим образом:

$$s(w) = s_{mch}(w) + c \cdot freq(w) \cdot len(w) - bndpenalty(w),$$

где s_{mch} — оценка процедуры MCHSR, которая соизмерима с суммой оценок символов из интервала отображения слова AP-цепи, $freq(w)$ — частота слова *w* в словаре, $len(w)$ — длина *w*, а $bndpenalty(w)$ — функция штрафа. Данная функция положительна, в случае если границы интервала отображения не являются символами-разделителями.

2. Отображение чисел, слов и последовательностей символов

Данные типы вершин одинаковы за исключением алфавита, поэтому без ограничения общности рассмотрим только числа. Число отображается на интервал, если все позиции внутри интервала имеют цифру в качестве альтернативы. Для каждого такого интервала ребро с типом «число» добавляется в хранилище ребер. В действительности в хранилище добавляются только позиции AP-цепи. Для того чтобы восстановить текстовое значение и оценку отображения используется так называемая редуцированная строка, т. е. строка с урезанным алфавитом. Редуцированная строка строится следующим образом. Пусть **A** некоторый алфавит, а AP-матрица состоит из альтернатив символов $\{c_i^k\}$, где i — номер знакоместа, а k — номер альтернативы. Определим строку $d_1 \ldots d_N$ следующей формулой:

$$d_i = \begin{cases} c_i^k & \text{if} \quad \exists k : c_i^k \in A, \forall t < k \;\; c_i^t \notin A, \\ * & \text{иначе,} \end{cases}$$

т. е. для каждого знакоместа выбирается альтернатива $c_i^k \in A$, имеющая наибольшую оценку. Она помещается на ту же позицию в редуцированной строке. Если не существует такой альтернативы, тогда символ '*' берется в качестве символа редуцированной строки, см. рис. 5. Также для каждой позиции строки сохраняется агрегированная оценка

$$s_+(n) = \sum_{i=1}^{n} s_i ,$$

s_i — оценка альтернативы d_i.

Оценка отображения впоследствии вычисляется как разность агрегированных оценок на краях интервала отображения, а текстовое значение соответствует подстроке редуцированной строки.

ЦБ-6.1580 от 08.04.98г.за
*6*6*1580*0**08*04*98**3*

Рис 5. Редуцированная строка с числовым алфавитом

3. Отображение PDS выражений

На данной стадии добавляются ребра соответствующие PDS выражениям. Для каждого интервала (i, j) мы пытаемся отобразить все PDS выражения на него. При отображении используется ряд оптимизационных методов и кэширование. PDS выражения обычно содержат от 2 до 6 вершин. Поэтому их отображение на AP-цепь осуществимо за приемлемое время. Если выражение отобразилось на некоторый интервал, в хранилище добавляется ребро с типом «PDS выражение». Текстовое значение ребра — конкатенация текстовых значений вершин выражения, а оценка отображения выражения — сумма оценок отображения вершин плюс некоторая премия пропорциональная длине интервала, на который отобразилось выражение. Если несколько выражений отобразилось на интервал, в хранилище добавляется только выражение получившее максимальную оценку. В хранилище добавляется тип ребра, ссылка на PDS выражение и путь через выражение. Путь определяется последовательностью вершин. На

рис. 4 ребро выражения «**от <#1–2> <month> <#2, 4> г.|года <->**» отображено на интервал [13–29]. Текст ребра — «21 июня 1990 г.».

4. Процедура поиска оптимального пути на АР-цепи

Процедура поиска оптимального пути основана на динамическом программировании. Вначале ребра наилучших альтернатив для каждого диапазона [i, j]. добавляются в хранилище. Текстовое значение у данных ребер — это наилучшие альтернативы АР-цепи в диапазоне [i, j]. Также интервал [i, j] обрабатывается общей процедурой, например, с использованием n-грамм либо словаря. После этого оптимальный путь находится с использованием техники динамического программирования. Оптимальный путь — это последовательность ребер различных типов. Ребра наилучших альтернатив также часто присутствуют в оптимальном пути. Они представляют фрагменты АР-цепи, которые не являются типичными для данного поля.

Объединение ребер оптимального пути образуют покрытие АР-цепи. Фрагмент оптимального пути, состоящий из отображенных ребер различных типов, показан на рис. 6.

5. Получение финального текстового значения

После того как оптимальный путь найден, финальное текстовое значение собирается из текстовых значений приписанных ребрам оптимального пути. Также на этом этапе выносится решение достоверности поля. Решение основано на надежности ребер.

PDS процедура оптимизирована для того, чтобы пост-обработка поля занимала допустимое время. Например, на компьютере с процессором Pentium IV 2 600 MHz в среднем поле «Назначение платежа» обрабатывается около 0,09 секунды. При этом средняя длина поля составляет 118 символов.

Список PDS выражений конструируется автоматически при помощи процедуры автоматического выделения синтаксиса (ASE процедуры). ASE процедура использует в качестве обучающей выборки образцы правильных текстовых значений обрабатываемого поля. Вначале вручную определяются используемые словари. Затем автоматически выделяется список ключевых слов. Ключевые слова — это наиболее часто встречаемые слова в обучающей выборке. После этого текстовые значения полей разбиваются по разделителям и исследуются n-ки слов. Если некоторое слово является словом из используемого словаря, то оно заменяется тэгом **<LEX_NODE>** (внутри тэга указан идентификатор словаря); если слово является числом, оно заменяется тэгом **<NUM_NODE>**. Также оце-

ниваются ограничения на длину. Преобразованные таким образом n-ки слов упорядочиваются по частоте деленной на **n**. Далее из полученного списка удаляются n-ки слов с частотой ниже некоторого порога. Затем производится сжатие списка путем слияния n-ок слов в PDS выражения. Две n-ки объединяются в PDS выражение с использованием операции ИЛИ, если они различаются только в одном слове. Сжатие полезно для компактного представления синтаксиса и как следствие этого для удобного дальнейшего редактирования синтаксиса, которое также может проводиться и вручную.

Рис 6. Оптимальный путь состоящий из ребер различного типа

4. Эксперименты

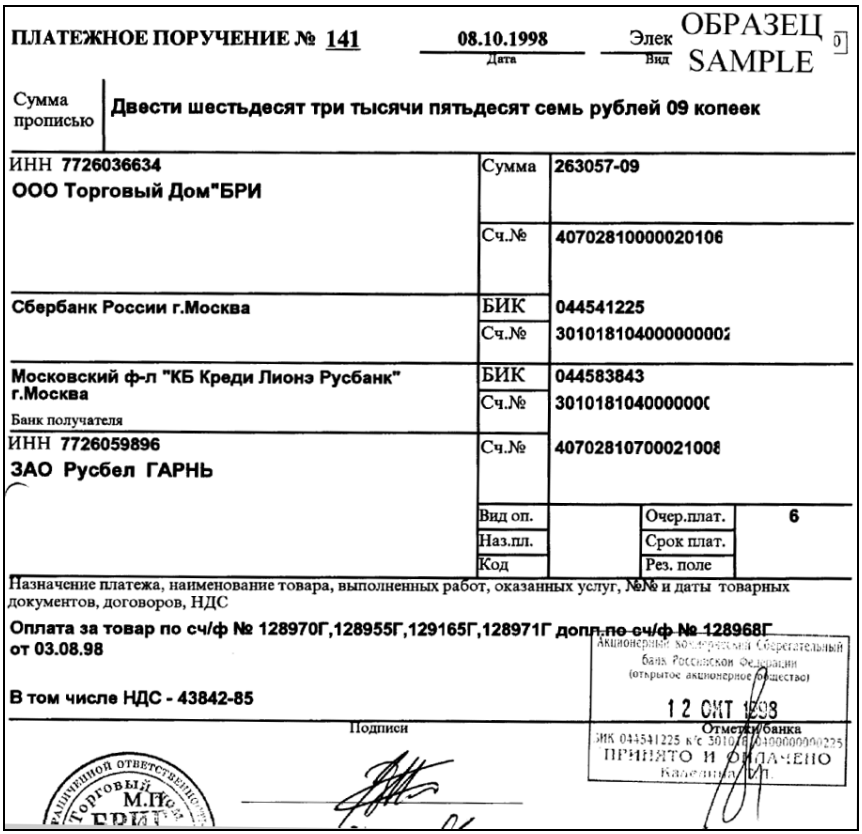
Для оценки качества работы PDS алгоритма были использованы платежные поручения ЦБРФ, содержащие ряд полей подходящих для постобработки, такие как «Сумма прописью», «Наименование плательщика/получателя», «Наименование банка плательщика/получателя», «Назначение платежа». Поле «Назначение платежа» допускает наиболее свободный синтаксис из полей указанных выше. Оно может содержать практически любой текст, объясняющий, за что был осуществлен данный платеж.

Но помимо этого Центральный Банк РФ предъявляет ряд требований к заполнению данного поля. Требуется, например, указать Налог на Добавленную Стоимость (НДС). Поэтому большинство полей содержат текстовую конструкцию типа «В т. ч. НДС-20 % — 1120-00» либо «НДС не облагается». Форма записи структуры может существенно меняться от поля к полю.

Как правило, платеж связан с некоторым договором и счетом фактурой. Поэтому текстовые структуры типа «По договору № 1267/99 от 18.01.99 г.» довольно часто присутствуют в данном поле. Также типичными для поля являются даты, временные периоды, имена и фамилии.

В процессе обучения процедура выделила 160 ключевых слов и 229 PDS выражений из 1 088 примеров текста содержащегося в поле «Назна-

чение платежа». Вначале были фиксированы используемые словари. Среди них словарь часто употребляемых Фамилий (5 434 фамилии), Имен (702 имени), Отчеств (2 547 отчеств), а также словарь Месяцев. Полученный PDS синтаксис покрывает 37 % текста в обучающей выборке. Средняя длина поля в ней составляет 118 символов. Список слов в обучающей выборки содержит 892 слова. Они были объединены в словарь и добавлены в список PDS выражений в качестве выражения состоящего из единственной словарной вершины. В некотором смысле это аналогично пост-обработке по специальному словарю фрагментов поля непокрытых PDS выражениями.



ПЛАТЕЖНОЕ ПОРУЧЕНИЕ № 141 08.10.1998 (Дата) Элек (Вид) ОБРАЗЕЦ SAMPLE 0

Сумма прописью: **Двести шестьдесят три тысячи пятьдесят семь рублей 09 копеек**

ИНН **7726036634**
ООО Торговый Дом"БРИ

Сумма **263057-09**

Сч.№ **40702810000020106**

Сбербанк России г.Москва

БИК **044541225**

Сч.№ **30101810400000000**

Московский ф-л "КБ Креди Лионэ Русбанк" г.Москва

Банк получателя

БИК **044583843**

Сч.№ **30101810400000000**

ИНН **7726059896**
ЗАО Русбел ГАРНЬ

Сч.№ **40702810700021008**

Вид оп. | Очер.плат. **6**
Наз.пл. | Срок плат.
Код | Рез. поле

Назначение платежа, наименование товара, выполненных работ, оказанных услуг, №№ и даты товарных документов, договоров, НДС

Оплата за товар по сч/ф № 128970Г,128955Г,129165Г,128971Г допл.по сч/ф № 128968Г от 03.08.98

В том числе НДС - 43842-85

Акционерный коммерческий Сберегательный банк Российской Федерации (открытое акционерное общество)
12 СЕНТ 1998
Отметки банка
БИК 044541225 к/с 30101810400000000225
ПРИНЯТО И ОПЛАЧЕНО

Подписи

Рис 7. Пример платежного поручения

Платежные поручения сканировались с разрешением 300 dpi и сохранялись в формате TIFF CCITT Group 4. Пример платежного поручения приведен на рис. 7. Тестовая выборка состояла из 758 платежных поручений отличных от документов из обучающей выборки.

Платежные поручения имеют гибкую геометрию с плавающими опорными элементами, отличающимися существенно от формы к форме. Локализация полей на формах такого рода задача тяжелая сама по себе. Иногда опорные линии могут отсутствовать и алгоритм, отображающий шаблон формы на изображения, в этом случае ориентируется на просветы между фрагментами текста [23–25]. Линии могут иметь разрывы или состоять из точек в случае печати на матричном принтере. Часто платежные поручения содержат печати и подписи внизу изображения, которые наползают на поле «Назначение платежа» (см. рис. 7). Также рядом с полем располагается статический текст, который требуется удалять в процессе распознавания. В результате перечисленных факторов, некоторые фрагменты поля могут быть утеряны на стадии локализации поля и впоследствии не могут быть восстановлены на стадии пост-обработки. Поэтому при расчете качества пост-обработки такого рода фрагменты учитываться не будут.

В табл. 1 приводится сравнение качества распознавания PDS алгоритма при использовании словаря состоящего из слов встреченных в процессе обучения (PDS + L), PDS алгоритма без использования данного словаря и результатов OCR распознавателя в чистом виде, т. е. без использования процедуры пост-обработки (CR).

Из таблицы видно, что наилучшее качество распознавания достигается при использовании PDS + L. При этом количество правильно распознанных полей увеличивается ~10 %. В то же время PDS + L дает более плохой результат в смысле неправильно распознанных и одновременно надежных полей, т. е. полей без сомнения (2,4 %). Качественная оценка PDS близка к PDS + L. Полей распознанных правильно и без сомнения больше у PDS + L, зато неправильных и без сомнения — меньше у PDS. Мы предпочли PDS + L т. к. он давал допустимое количество ошибок, а количество правильных и надежных полей повышал существенно (примерно на 6 %).

Для сравнения PDS подхода (PDSA) с Синтаксическим Подходом (SA) мы использовали поле «Почтовый адрес» на анкетах Пенсионного Фонда РФ (форма АДВ1). Данное поле имеет довольно свободный синтаксис, но в то же время оно полностью может быть описано синтаксическими диаграммами.

Таблица 1

Результаты PDS алгоритма полученные на поле «Назначение платежа»

S (758 полей)	PDS + L	PDS	CR
всего форм		758 (100 %)	
непривязанных форм		14 (1,8 %)	
поля:			
полей распознано		709 (93,5 %)	
правильно	503 (71 %)	489 (69 %)	432 (61 %)
правильно / надежно	361 (51 %)	319 (45 %)	174 (25 %)
правильно / не надежно	142 (20 %)	170 (24 %)	258 (36 %)
неправильно	206 (29 %)	220 (31 %)	277 (39 %)
неправильно / надежно	17 (2,4 %)	13 (1,8 %)	7 (1,0 %)
неправильно / не надежно	189 (27 %)	207 (29 %)	270 (38 %)
символы:			
правильно локализованные		83 %	
неправильно локализованные		17 %	
правильно распознанные	98,1 %	97,6 %	94,3 %
неправильно распознанные	1,9 %	2,4 %	5,7 %

Подход был проверен на анкетах предоставленных Пенсионным фондом РФ. Анкеты сканировались с разрешением 200 dpi и сохранялись в формате TIFF CCITT Group 4. Форма АДВ1 разработана для рукопечатного заполнения. Для описания допустимого синтаксиса адреса была использована БД КЛАДР Налоговой Службы РФ. Также было проанализировано порядка 15 тысяч адресов реально присутствующих на форме АДВ1.

Множество из 3 674 форм АДВ1 было разделено на две части — обучающая выборка S_1 состоящая из 1 260 форм и тестовая выборка S_2 из 2 414 форм. В обучающей выборке оказалось 1 387 адресов, а в тестовой — 2 718 адресов. Каждая форма содержала от одного до двух адресов — адрес по прописке и адрес проживания.

Для построения синтаксических диаграмм (СД) была использована полуавтоматическая процедура определения ограничений на размер и положение границ вершин СД. Полученная СД адреса представляет собой довольно большой граф, состоящий из 45 вершин и пяти вложенных поддиаграмм. Также СД содержит ряд тяжеловесных словарных вершин, например Улицы (49 042 элементов), Населенные пункты — 11 449 элементов и Города — 1 156 элементов. Все словари содержат слова в морфологически нормализованной форме. Для отображения словарных вершин использовался специальный алгоритм MCHSR быстрого отображения словаря на AP-цепь.

Таблица 2

Экспериментальные результаты пост-обработки почтового адреса

S_1 (1 387 адресов)	**PDSA**	**SA**	**CR**
Распознано		1 375	
Отбраковано		12	
Правильных	84,9 %	89,7 %	54,2 %
правильных / надежно	56,8 %	67,0 %	21,1 %
правильных / не надежно	28,1 %	22,7 %	33,1 %
неправильных	15,1 %	10,3 %	45,8 %
неправильных / не надежно	14,0 %	8,8 %	34,6 %
неправильных / надежно	1,1 %	1,5 %	11,2 %
S_2 (2 718 адресов)	**PDSA**	**SA**	**CR**
Распознано		2 718	
отбраковано		29	
правильных	84,2 %	88,1 %	57,8 %
правильных / надежно	53,8 %	62,9 %	26,5 %
правильных / не надежно	30,4 %	25,2 %	31,3 %
неправильных	15,8 %	11,9 %	42,2 %
неправильных / не надежно	14,5 %	10,0 %	29,1 %
неправильных / надежно	1,3 %	1,9 %	13,1 %

Для получения списка PDS выражений была использована ASE процедура. При этом в качестве входных данных использовалось множество из 15 000 адресов.

Полученные результаты приведены в табл. 2. Столбцы PDSA, SA и CR таблицы относятся к PDS подходу, Синтаксическому подходу и к распознаванию без использования пост-обработки соответственно.

Как показано в табл. 2, Синтаксический подход имеет более высокое качество, чем PDS подход, но не сильно. При этом разработка и настройка процедуры пост-обработки адреса с использованием PDS заняла значительно меньше времени, чем при использовании Синтаксического подхода, которая включала построение СД и тонкую ручную настройку функций отображения вершин СД на АР-цепь. Так что применение PDS подхода допустимо и оправдано в случае недостатка времени.

5. Реализация, применения и выводы

PDS подход был реализован в системе массового ввода документов Cognitive Forms [21] и был успешно применен при обработке платежных документов ЦБРФ и анкет Пенсионного Фонда РФ. На данный момент системой Cognitive Forms при использовании PDS подхода было обработано более 10 миллионов документов. Качество PDS подхода измерялось на поле «Назначение платежа» платежных поручений. Данное поле содержит много пунктуации, сокращений и аббревиатур. Также довольно много документов напечатано на матричном принтере. Поэтому качество распознавания без использования пост-обработки не высоко. На этапе обучения ASE процедура выделила 160 ключевых слов и 229 PDS выражений из более 1 000 образцов текста. Полученный частичный синтаксис покрывает около 40 % текста. Также был построен полный список слов содержащихся в обучающей выборке. Данный список использовался в качестве PDS выражения состоящего лишь из одной словарной вершины. PDS подход повысил качество распознавания символов с 94 % до 98 %.

Также мы сравнивали PDS подход (PDSA) с Синтаксическим подходом (SA) на примере пост-обработки рукопечатного почтового адреса. Синтаксический подход дал примерно 90 % правильно распознанных полей, а PDSA — примерно 85 %.

При использовании SA, адреса, не удовлетворяющие заданному нами синтаксису часто (примерно в 2 % случаев) ошибочно укладывались в

синтаксическую диаграмму. Это приводило к некоторому замешательству среди операторов ввода. Но что касается PDS подхода, то он производил замены в более осторожной манере.

Если настройка PDS алгоритма для пост-обработки почтового адреса заняла у нас несколько дней, то описание синтаксиса адреса заняла у нас более месяца. Так что PDSA — метод гораздо более быстрый в смысле разработки функций пост-обработки определенных полей. Но наряду с этим PDSA не позволяет структурировать и нормализовать финальное текстовое значение, что часто необходимо.

Какой подход использовать зависит от специфики задачи. Если требуется быстро разработать алгоритм с довольно хорошим качеством распознавания — следует использовать PDSA. Если же напротив, требования к качеству распознавания чрезвычайно высоки и имеется достаточно времени, лучше использовать SA, который является более гибким, а также позволяет использовать внешнюю информацию.

Благодарности

Авторы хотят поблагодарить коллег из группы научных исследований Cognitive Technologies за полезные советы и участие, к. ф. н. Елену Борисовну Козеренко за реферирование статьи, Пенсионный Фонд РФ за предоставленные материалы и данные. Также в особенности хотелось поблагодарить член.-корр. РАН Владимира Львовича Арлазарова за профессиональную помощь и участие в проведении и реализации данного проекта.

Литература

1. *Sargur Srihari N.* Document image understanding // Proc. of 1986 fall joint computer conference on Fall joint computer conference, November 1997. P. 87–96.
2. *Kukich K.* Techniques for automatically Correcting Words in Text // ACM computing survey Computational Linguistic. V. 24. № 4. P. 377–439, 1992.
3. *Mailburg Michael H.* Comparative Evaluation of Techniques for Word Recognition Improvement by Incorporation of Syntactic Information // 4th International Conference Document Analysis and Recognition (ICDAR ‘97) August 1997. P. 784.
4. *Beitzel S., Jensen E., Grossman D.* A Survey of Retrieval Strategies for OCR Text Collections // Proc. of 2003 Symposium on Document Image Understanding Technology, April 2003.

5. *Sholomov D. L.* Syntactical Approach to Post-Processing of Fuzzy recognized Text // Proc. of The International Conference on Machine Learning, Technologies and Applications, CSREA Press. P. 115–121. June 2003, USA.
6. *Sholomov D. L.* Interpreting the Indistinctly Recognized Textual Constructions // Pattern Recognition and Image Analysis. 2003. V. 13. № 2. P. 353–355.
7. *Brakensiek A., Rottland J., Rigoll G.* Handwritten Address Recognition with Open Vocabulary Using Character N-grams // Proc. of 8th International Workshop on Frontiers in Handwriting Recognition (IWFHR), 2002.
8. *Chen D. Y., Mao J., Mohiuddin K.* An Efficient Algorithm for Matching a Lexicon with a Segmentation Graph // Fifth International Conference on Document Analysis and Recognition, India, September 1999.
9. *Bouchaffra Djamel and Govindaraju Venu and Srihari Sargur N.* Postprocessing of Recognized Strings Using Nonstationary Markovian Models // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997. V. 21. № 10. P. 990–999.
10. *Brown P. F., Della Pietra V. J., de Souza P. V., Lai J. C., Mercer R. L.* Class-Based n-gram Models of Natural Language. // Computational Linguistics. 1992. V. 18. № 4. P. 467–479.
11. *Niyogi D., Srihari S. N., and Govindaraju V.* Analysis of printed forms // H. Bunke and P. S. P. Wang, editors, Handbook on Optical Character Recognition and Document Image Analysis. World Scientific Publishing Co., Singapore, 1996.
12. *Jain A. K., Duin R. P. W., Mao J.* Statistical Pattern Recognition: A Review // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002. V. 22. № 1. P. 4–37.
13. *Aho A., Sethi R., Ullman J.* Compilers: principles, techniques and tools. N. Y.: Addison-Wesley, 1986.
14. *Blumenstein M., Verma B.* A Neural Network for Real-World Postal Address Recognition.
15. *Wong P. K., Ho T. K., Srihari S. N.* Firm Name Recognition for Automatic Address Interpretation. // Proc. of the 5th {USPS} Advanced Technology Conference, November 1992. P. 757–770.
16. *Srihari Sargur and Shin Yong-Chul and Ramanaprasad Vemulapati and Lee Dar-Shyang*. A System to Read Names and Addresses on Tax Forms.
17. *Ho T. K. and Hull J. J. and Srihari S. N.* Word Recognition with Multi-Level Contextual Knowledge // Proc. of the 1st Int'l Conference on Document Analysis and Recognition, October 1991. P. 905–915.
18. *Schuermann J.* A Multifont Word Recognition System for Postal Address Reading // IEEE Transactions on Computers, C-27, 8, August 1978, 721–732. 9.
19. *Almendra Freitas C. O. de, El Yacoubi A., Bortolozzi F., Sabourin R.* Brazilian Bank Check Handwritten Legal Amount Recognition // Proc. of the XIII Brazilian Symposium on Computer Graphics and Image Processing.

20. *Арлазаров В. Л., Славин О. А.* Алгоритмы распознавания и технологии ввода текстов в ЭВМ // Информационные технологии и вычислительные системы 1996. № 1, 6. С. 48–54.
21. *Арлазаров В. В, Постников В. В., Шоломов Д. Л.* Cognitive Forms — система массового ввода структурированных документов // Сборник «Управление информационными потоками». М.: УРСС, 2002. С. 37–49.
22. *Misyurev A. V.* Hand-Printed Character Recognition by Neural Networks // Proc. of the 5th German-Russian Workshop on Pattern Recognition and Image Understanding (GRWS98), 1999.
23. *Postnikov V. V.* Flexible forms identification // Proc. of the 5th German-Russian Workshop on Pattern Recognition and Image Understanding (GRWS98), 1999.
24. *Постников В. В.* Автоматическая идентификация и распознавание структурированных документов // Дис. … канд. техн. наук (спец. 05.13.01). М., 2001.