MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 30, NO. 4, 2023
journal homepage: www.mais-journal.ru

ARTIFICIAL INTELLIGENCE

Extracting named entities from Russian-language documents with different expressiveness of structure

M. D. Averina[1], O. A. Levanova[1]

DOI: 10.18255/1818-1015-2023-4-382-393

[1]P.G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.

MSC2020: 68T50
Research article
Full text in Russian

Received October 13, 2023
After revision November 10, 2023
Accepted November 15, 2023

This work is devoted to solving the problem of recognizing named entities for Russian-language texts based on the CRF model. Two sets of data were considered: documents on refinancing with a good document structure, semi-structured texts of court records. The model was tested under various sets of text features and CRF parameters (optimization algorithms). In average for all entities, the best F-measure value for structured documents was 0.99, and for semi-structured ones 0.86.

Keywords: named entity extraction; CRF

INFORMATION ABOUT THE AUTHORS

Maria D. Averina
corresponding author
orcid.org/0009-0005-3111-1488. E-mail: maverina518@gmail.com
graduate student.

Olga A. Levanova
orcid.org/0000-0001-8078-4447. E-mail: olaydy@gmail.com
PhD, Associate professor.

For citation: M. D. Averina and O. A. Levanova, “Extracting named entities from Russian-language documents with different expressiveness of structure”, *Modeling and analysis of information systems*, vol. 30, no. 4, pp. 382-393, 2023.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 30, № 4, 2023
сайт журнала: www.mais-journal.ru

Извлечение именованных сущностей из русскоязычных документов с различной выраженностью структуры

М. Д. Аверина[1], О. А. Леванова[1]

DOI: 10.18255/1818-1015-2023-4-382-393

[1]Ярославский государственный университет им. П.Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 13 октября 2023 г.
После доработки 10 ноября 2023 г.
Принята к публикации 15 ноября 2023 г.

Данная работа посвящена решению задачи распознавания именованных сущностей для русскоязычных текстов на основе модели CRF. Рассмотрены два набора данных: документы о рефинансировании с хорошей структурой документа, слабоструктурированные тексты судебных протоколов. Было проведено тестирование модели при различных наборах текстовых признаков и параметрах CRF (алгоритмов оптимизации). В среднем по всем сущностям лучшее значение F-меры для структурированных документов составило 0.99, а для слабоструктурированных 0.86.

Ключевые слова: извлечение именованных сущностей; CRF

ИНФОРМАЦИЯ ОБ АВТОРАХ

Мария Дмитриевна Аверина автор для корреспонденции	orcid.org/0009-0005-3111-1488. E-mail: maverina518@gmail.com аспирант.
Ольга Александровна Леванова	orcid.org/0000-0001-8078-4447. E-mail: olaydy@gmail.com канд. физ.-мат. наук, доцент.

Для цитирования: M. D. Averina and O. A. Levanova, “Extracting named entities from Russian-language documents with different expressiveness of structure”, *Modeling and analysis of information systems*, vol. 30, no. 4, pp. 382-393, 2023.

Введение

В современную эпоху цифровизации всё большую популярность обретают технологии электронного документооборота. В этой связи усиливается интерес к автоматизации обработки электронных документов путём извлечения из них информации. Например, бывает полезно выделить все даты и автоматически отсортировать документы. Или же, после сканирования банковских документов удобно автоматически определить не только ФИО клиента, но и имя сотрудника, паспортные данные клиента или сумму кредита, такая информация может существенно упростить процесс обработки электронных документов.

Задача распознавания именованных сущностей (NER) состоит в автоматическом выявлении текстовых фрагментов, которые несут определенную смысловую нагрузку. Например, в классической задаче выделены следующие классы: человек (PER), местоположение (LOC), организация (ORG) и другие (MISC) [1]. Для других предметных областей сущности могут быть менее обобщенными и более сложными. Так, для судебных протоколов может быть полезно выделить не просто все фамилии в документе, а более конкретно — ФИО истца и судьи. Примером сложной сущности может быть решение суда, которое «размазано» по тексту (сущности соответствует не сплошной фрагмент текста).

Данная статья рассматривает решение задачи распознавания именованных сущностей для русскоязычных судебных и банковских документов. Тексты в таких документах отличаются от текстов в других предметных областях и не похожи между собой. Стоит отметить, что банковские документы имеют стандартную структуру, а судебные протоколы написаны в более свободном стиле и плохо структурированы.

Задача извлечения именованных сущностей особенно актуальна для русского языка, поскольку почти все существующие системы и библиотеки хорошо работают с английскими текстами, а для других языков результаты [2] значительно хуже (или же их вообще нет). Статей посвященных задаче NER для не классических сущностей крайне мало. Наибольший интерес представляет статья языке для новостных текстов на казахском [3], в которой ищется 25 сущностей.

Классическим подходом для решения поставленной задачи является модель CRF [4]. Однако ее качество очень сильно зависит от предварительной обработки текста и внутренних параметров алгоритма. В статье исследованы различные текстовые признаки и параметры модели CRF.

1. Наборы данных

Для данного исследования были использованы два набора данных: судебные протоколы и договоры о консолидации и рефинансировании задолженности. Рассмотрим каждый тип документа подробнее.

1.1. Судебные протоколы

Из-за специфики задачи в документах были выделены сущности, которые характерны для юридической области. Так, например, из текста необходимо было выделить не просто ФИО всех людей, но также дифференцировать их: ответчик это, истец или судья. Данные документы были взяты из открытой базы судебной статистики (http://www.cdep.ru), которая содержит анонимизированные судебные решения. Такой выбор был сделан из-за большого количества имеющихся документов и характера текстов, а именно: тексты содержат множество различных имён, дат, номиналов и сумм. Стоит отметить, что данные документы практически не структурированы, что усложняет решение поставленной задачи.

№ <*doc num, 2-1606/2018*>

РЕШЕНИЕ
Именем Российской федерации
<*date court, 02 ноября 2018 года*> <*court, Кировский районный суд г.Томска*> в составе:
председательствующего судьи <*judge, Алиткиной Т.А.*>,
при секретаре Бондаревой Е.Е.,
с участием представителя процессуального истца старшего прокурора Кировского района г.Томска Морарь И.В., материального истца <*plaintiff, Полищука Э.Г.*>, представителя ответчика <*defendant, Семенова С.М*>., действующего на основании доверенности от 17.05.2017 (срок действия доверенности три года),

...

собственному желанию, взыскании задолженности по заработной плате, компенсации, предусмотренной ст.236 ТК РФ, компенсации за задержку выплаты заработной платы, взыскании денежной компенсации морального вреда <*court decision, удовлетворить частично*>.

Fig. 1. Example of a court record. Beginning of text.

Рис. 1. Пример судебного протокола. Начало текста.

<*payment fine, задолженность по заработной плате*> за период с 26.12.2017 по 22.01.2018 в размере <*payment amount, 20 520,00 руб.*>; <*payment fine,денежную компенсацию, предусмотренную ст.236 ТК РФ*>, в размере <*payment amount,796,20 руб.*>; <*payment fine,утраченный заработок за время вынужденного прогула*> в размере <*payment amount, 91 198,80 руб.*>, денежную <*payment fine,компенсацию морального вреда* >в размере <*payment amount, 3 000 руб*>.

...

Решение может быть обжаловано в Томский областной суд путем подачи апелляционной жалобы через Кировский районный суд г.Томска <*appeal time, в течение 1 месяца со дня изготовления решения в мотивированном виде*>.

Судья: (подпись) <*judge, Т.А.Алиткина*>

Fig. 2. Example of a court record. End of text.

Рис. 2. Пример судебного протокола. Конец текста.

На основе данной базы была сформирована выборка из 344 файлов (документы состоят из 4–7 страниц), и размечена с помощью инструмента BRAT (онлайн-инструмент для разметки письменных текстов https://brat.nlplab.org). На рисунках 1 и 2 показан пример разметки документа. Все сущности условно можно разделить на группы, характеризующие сложность их распознавания. В группу простых сущностей входят номер документа (doc num), решение суда, судья и другие (рис. 1). Встречаются и сложные сущности, которые содержат более двух слов, такие как время обжалования, истец, штраф, сумма платежа, ответчик и суд.

Среднее количество слов для каждой сущности и процент документов, содержащих соответствующие объекты представлены в таблице 1. Некоторые сущности встречаются в меньшем числе документов, что может сказаться на качестве распознавания.

Более того, некоторые сущности вариативны и зависят от содержания документа. Например, ответчиком может быть физическое лицо, организация или представитель ответчика. Уплата штрафа

Table 1. Entities analysis for court records

Таблица 1. Анализ сущностей судебных протоколов

Сущность	Размер сущности	Процент встречаемости в документах
истец	2.4	98
статья или тип штрафа	2.4	77
сумма выплаты	8.2	70
судья	1.9	98
номер документа	0.9	78
ответчик или представитель	6.0	98
дата суда	2.6	93
суд	4.2	98
решение суда	1.9	94
срок обжалования	9.6	95

Table 2. Entities analysis of refinancing agreements

Таблица 2. Анализ сущностей для договоров о рефинансировании

Сущность	Размер сущности	Процент встречаемости в документах
клиент 1	3.0	100.0
номер паспорта	1.0	63.4
номер документа	3.5	99.0
число выплат	1.0	90.0
клиент 2	6.0	100.0
сумма прощения	2.0	100.0
дата документа	1.5	97.0
день рождение	1.6	77.0
месячная выплата	2.0	99.0
номер кредитного договора	1.0	100.0
имя банка	3.0	100.0
дата кредитного договора	1.0	98.5
итоговая сумма	4.0	100.0
дата первого платежа	1.0	93.0
день оплаты	4.0	96.0
адрес клиента	7.0	63.0

в разных документах может иметь различные значения, такие как статья закона, платеж или компенсация. Сложные сущности могут быть прерывистыми, например штраф и сумма платежа (рис. 2), и поэтому их труднее распознать.

1.2. Договоры о консолидации и рефинансировании задолженности

Второй набор данных был предоставлен одной из коммерческих организаций, заинтересованной в автоматизации процессов документооборота. Данные договоры можно отнести к полуструктурированному типу, поскольку часть сущностей располагается в таблицах, а другие в сплошном

тексте. Стоит отметить, что структура документа у различных организаций отличается, поэтому результаты выделения именованных сущностей могут отличаться.

Экспертами было размечено 200 договоров о рефинансировании при помощи инструмента BRAT. В выборке преобладают документы состоящие из одной страницы, но также встречаются и двухстраничные. Пример такого документа представлен на рисунке 3 (документ изменен в целях конфиденциальности информации). Для поставленной задачи требовалось выделить 16 сущностей, в основном простых. Из таблицы 2 видно, что практически все сущности содержатся в каждом документе, кроме номера паспорта и адреса клиента.

2. Извлечение признаков

Первым этапом работы с текстом является его предобработка: токенизация, удаление лишних слов (например, стоп-слов) или символов. Затем необходимо извлечь признаки из текста. Существует несколько подходов: основанные на регулярных выражениях, морфологических признаках, синтаксическом и семантическом анализе. Наиболее очевидным и простым решением является извлечение признаков при помощи регулярных выражений. Например, можно извлечь информацию о ближайших знаках препинания или регистре букв: номер документа «№ 11255588» делится на «№» и «11255588», где № идентифицируется как специальный символ. В работе были использованы следующие специальные символы: @, #, №, \, %, $, |.

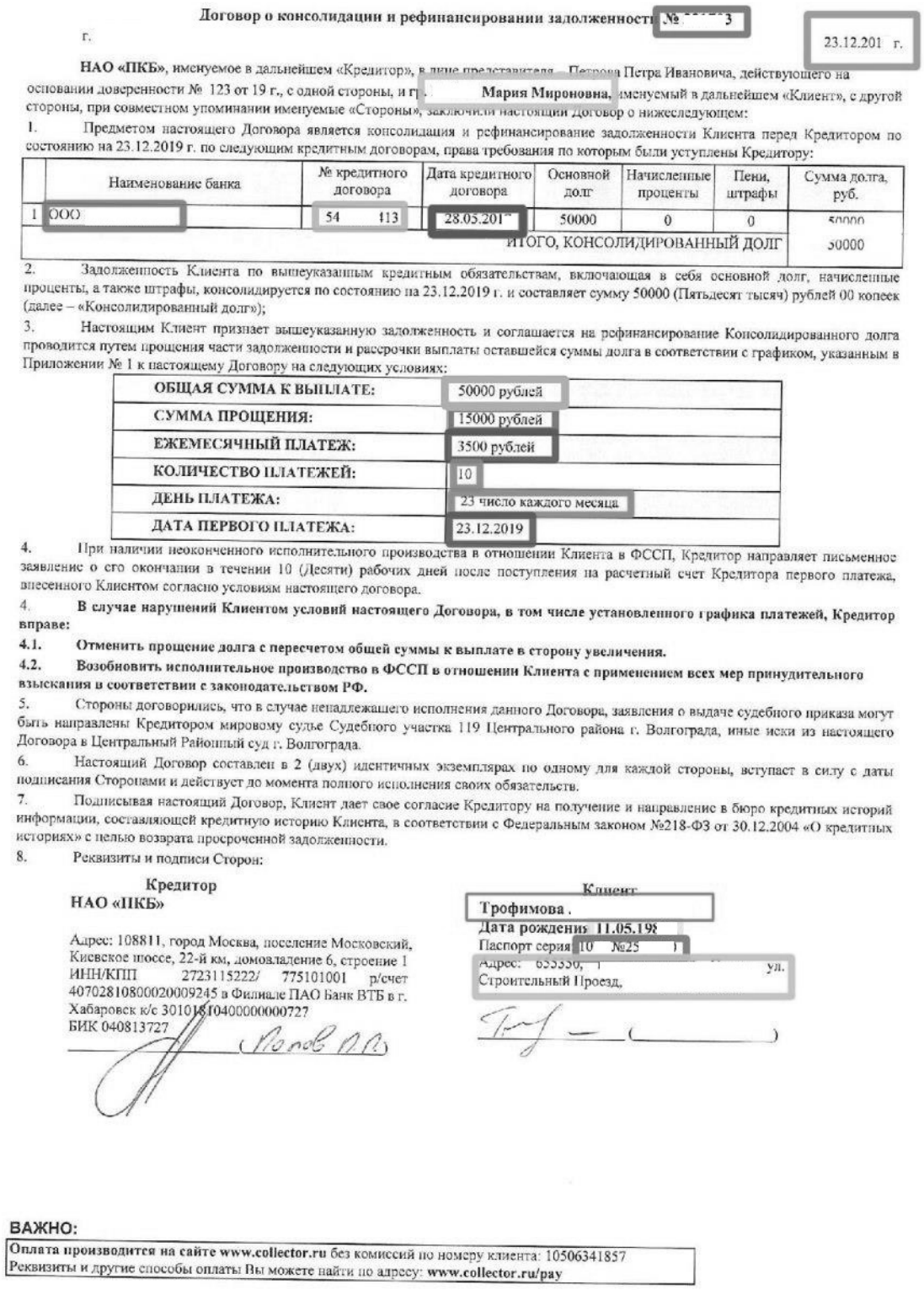
Ниже представлен список признаков, основанных на регулярных выражениях:

- первая буква прописная;
- первая буква маленькая;
- все буквы маленькие;
- все буквы заглавные;
- наличие @ внутри слова;
- наличие запятой и (или) точки в конце (начале) слова;
- есть ли в слове цифры.

Все слова в тексте имеют различные падежи и склонения, что в свою очередь может усложнять работу модели. Одним из способов решения данной проблемы является нормализация — приведение слова в начальную форму при помощи лемматизации или стемминга. При этом для каждого слова сохраняются признаки такие как: число и род и т. д. Стоит отметить, что часть речи также является морфологическим признаком, и в случае слов-спецсимволов каждому символу присваивается уникальное значение части речи.

Так же в качестве признака можно использовать само «слово», но данный признак не является информативным из-за большой вариативности. Например, если в используемых документах часто фигурирует фамилия судьи и в выборке много дел с одним судьей, то на других данных (с другими судьями) фамилия судьи находиться не будет.

Следующим по популярности подходом к вычислению признаков является векторизация слов — представление слов в виде вектора чисел. Для данного подхода были выбраны алгоритмы Word2Vec [5] и FastText [6], основанные на контекстной близости слов. Алгоритм Word2Vec работает с большими текстовыми данными и по определенным правилам, присваивает каждому слову уникальный набор чисел — семантический вектор. Спустя время были представлены улучшенные модификации данного алгоритма, одной из них является FastText. FastText исправляет недостаток Word2Vec, заключающийся в невозможности представления слова в виде вектора, если его не было в обучающем наборе.



Договор о консолидации и рефинансировании задолженности № 3

г. 23.12.201 г.

НАО «ПКБ», именуемое в дальнейшем «Кредитор», в лице представителя Петрова Петра Ивановича, действующего на основании доверенности № 123 от 19 г., с одной стороны, и г **Мария Мироновна**, именуемый в дальнейшем «Клиент», с другой стороны, при совместном упоминании именуемые «Стороны», заключили настоящий Договор о нижеследующем:

1. Предметом настоящего Договора является консолидация и рефинансирование задолженности Клиента перед Кредитором по состоянию на 23.12.2019 г. по следующим кредитным договорам, права требования по которым были уступлены Кредитору:

	Наименование банка	№ кредитного договора	Дата кредитного договора	Основной долг	Начисленные проценты	Пени, штрафы	Сумма долга, руб.
1	ООО	54 113	28.05.201	50000	0	0	50000
				ИТОГО, КОНСОЛИДИРОВАННЫЙ ДОЛГ			50000

2. Задолженность Клиента по вышеуказанным кредитным обязательствам, включающая в себя основной долг, начисленные проценты, а также штрафы, консолидируется по состоянию на 23.12.2019 г. и составляет сумму 50000 (Пятьдесят тысяч) рублей 00 копеек (далее – «Консолидированный долг»);

3. Настоящим Клиент признает вышеуказанную задолженность и соглашается на рефинансирование Консолидированного долга проводится путем прощения части задолженности и рассрочки выплаты оставшейся суммы долга в соответствии с графиком, указанным в Приложении № 1 к настоящему Договору на следующих условиях:

ОБЩАЯ СУММА К ВЫПЛАТЕ:	50000 рублей
СУММА ПРОЩЕНИЯ:	15000 рублей
ЕЖЕМЕСЯЧНЫЙ ПЛАТЕЖ:	3500 рублей
КОЛИЧЕСТВО ПЛАТЕЖЕЙ:	10
ДЕНЬ ПЛАТЕЖА:	23 число каждого месяца
ДАТА ПЕРВОГО ПЛАТЕЖА:	23.12.2019

4. При наличии неоконченного исполнительного производства в отношении Клиента в ФССП, Кредитор направляет письменное заявление о его окончании в течении 10 (Десяти) рабочих дней после поступления на расчетный счет Кредитора первого платежа, внесенного Клиентом согласно условиям настоящего договора.

4. **В случае нарушений Клиентом условий настоящего Договора, в том числе установленного графика платежей, Кредитор вправе:**

4.1. Отменить прощение долга с пересчетом общей суммы к выплате в сторону увеличения.

4.2. Возобновить исполнительное производство в ФССП в отношении Клиента с применением всех мер принудительного взыскания в соответствии с законодательством РФ.

5. Стороны договорились, что в случае ненадлежащего исполнения данного Договора, заявления о выдаче судебного приказа могут быть направлены Кредитором мировому судье Судебного участка 119 Центрального района г. Волгограда, иные иски из настоящего Договора в Центральный Районный суд г. Волгограда.

6. Настоящий Договор составлен в 2 (двух) идентичных экземплярах по одному для каждой стороны, вступает в силу с даты подписания Сторонами и действует до момента полного исполнения своих обязательств.

7. Подписывая настоящий Договор, Клиент дает свое согласие Кредитору на получение и направление в бюро кредитных историй информации, составляющей кредитную историю Клиента, в соответствии с Федеральным законом №218-ФЗ от 30.12.2004 «О кредитных историях» с целью возврата просроченной задолженности.

8. Реквизиты и подписи Сторон:

Кредитор
НАО «ПКБ»

Адрес: 108811, город Москва, поселение Московский, Киевское шоссе, 22-й км, домовладение 6, строение 1
ИНН/КПП 2723115222/ 775101001 р/счет 40702810800020009245 в Филиале ПАО Банк ВТБ в г. Хабаровск к/с 30101810400000000727
БИК 040813727

Петров П.П.

Клиент
Трофимова .
Дата рождени 11.05.198
Паспорт серия 10 №25
Адрес: 655550, г ул. Строительный Проезд,

( )

ВАЖНО:

Оплата производится на сайте www.collector.ru без комиссий по номеру клиента: 10506341857
Реквизиты и другие способы оплаты Вы можете найти по адресу: **www.collector.ru/pay**

Fig. 3. Example of a refinancing agreement.

Рис. 3. Пример договора о рефинансировании.

В дальнейшем мы будем использовать следующие обозначения признаков: «регулярные выражения» — *r*, «само слово» — *v*, часть речи — *m*, нормализация — *n*, Word2Vec — *w*, FastText — *f*. Для повышения качества распознавания сущностей полезно учитывать контекст слова в тексте. Таким образом, в качестве характеристик слова мы также использовали признаки его соседей. Цифра после буквы обозначает количество рассматриваемых соседей влево и вправо. Например, *f3* означает, что использовалось 7 признаков FastText — для текущего слова и для 3-х соседей с каждой стороны.

3. Метод CRF

Как уже было сказано, наиболее популярным и традиционным подходом для решения задачи распознавания именованных сущностей является метод conditional random field (CRF) [7, 8]. Данный алгоритм оптимизирует всю цепочку меток целиком, а не каждый элемент по отдельности, учитывая особенности и взаимозависимости в данных. Поэтому данная модель хорошо подходит для решения задач сегментации и маркировки последовательностей. Основная идея CRF заключается в том, чтобы предсказать последовательность меток или классов для входной последовательности на основе набора наблюдаемых признаков. У CRF есть несколько ключевых особенностей:

1. Условная зависимость: CRF моделирует условные вероятности меток в зависимости от признаков. Модель учитывает контекст и взаимодействие между соседними элементами входной последовательности при принятии решения о метках.
2. Марковские свойства: CRF основывается на марковских свойствах, что означает, что вероятность текущей метки зависит только от предыдущих меток в последовательности. В контексте последовательностей, CRF моделирует условные вероятности меток, основываясь на предыдущих метках, а также на наблюдаемых признаках.
3. Графическая структура: CRF представляет собой графическую модель, в которой узлы соответствуют элементам входной последовательности, а ребра представляют зависимости между ними. Эта структура позволяет моделировать сложные зависимости между метками.

Рассмотрим модель подробнее. Пусть у нас есть наблюдаемые переменные (входные признаки) $X = x_0, ..., x_T$ и скрытые переменные (метки) $Y = y_0, ..., y_T$, где T - длина последовательности. CRF моделирует условные вероятности $P(Y|X)$ с использованием следующей формулы:

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_k \sum_i \lambda_k f_k(y_{i-1}, y_i, x_i)\right),$$

где $Z(X)$ — нормализующий множитель, λ_k — параметры модели, f_k — функции признаков.

Нормализующий множитель гарантирует, что сумма всех возможных состояний меток в последовательности будет равна 1. Функции признаков f_k могут быть заданы различными способами, и выбор конкретных функций зависит от предметной области и структуры данных. Они могут учитывать контекстную информацию, зависимости между соседними элементами, а также другие свойства данных.

Обучение CRF модели включает настройку параметров λ_k с использованием метода максимального правдоподобия. Для нахождения оптимальных параметров модели часто применяются методы оптимизации, такие как градиентный спуск или условный градиентный спуск.

В данной работе была использована реализация CRF из библиотеки sklearn crfsuite. Ниже представлен список поддерживаемых оптимизаторов:

- *l2sgd* — стохастический градиентный спуск с регуляризацей *L2*;
- *lbfgs* — градиентный спуск с использованием алгоритма Бройдена-Флетчера-Гольдфарба-Шанно (*L1* и *L2*);

- *ap* — усредненный персептрон (Averaged Perceptron);
- *pa* — passive aggressive, алгоритм, основанный на бинарной классификации [9];
- *arow* — метод адаптивной регуляризации весов вектора.

4. Тестирование

Перечислим сначала библиотеки, которые использовались при реализации: NLTK [10], gensim [11], pymorphy2 [12], sklearn crfsuite (http://www.chokkan.org/software/crfsuite/).

Прежде чем перейти к тестированию, стоит обсудить оценку качества распознавания. Существует два подхода к оценке качества решения задачи NER: F1-мера, предложенная на конференции CoNLL [13], которая учитывает всю цепочку слов сущности; и стандартная F1-мера для каждой сущности по отдельности.

Поскольку в наших данных есть разрывные сущности, было решено использовать стандартную F1-меру для каждой сущности по отдельности. Данный подход дает несколько более оптимистичные результаты, однако тестирование показало, что метрики не сильно отличаются на наших наборах данных (0.04 ± 0.02).

4.1. Результаты на документах судебной статистики

Обучение проводилось на 241 судебном протоколе, а отложенная выборка состояла из 103 документов. В данной статье авторами были протестированы различные наборы признаков и алгоритмов оптимизации. В таблице 3 приведен анализ оптимизаторов, при фиксированных наборах признаков.

Table 3. Quality analysis. Court records.

Таблица 3. Анализ качества распознавания. Судебные протоколы.

Сущность	r1, v1 pa	r3, v3 pa	r3, v3, m3 pa	f3(10), r3, m3 lbfgs
срок обжалования	0.92	0.93	0.93	0.92
суд	0.94	0.97	0.97	0.91
решение суда	0.71	0.71	0.71	0.60
дата суда	0.89	0.94	0.92	0.82
ответчик или представитель	0.52	0.61	0.67	0.57
номер документа	0.95	0.98	0.97	0.95
судья	0.97	0.97	0.97	0.95
сумма выплаты	0.64	0.73	0.77	0.72
статья или тип штрафа	0.64	0.69	0.68	0.55
истец	0.85	0.86	0.86	0.82
F-macro	0.82	0.85	0.86	0.80

Следует отметить, что изменение количества соседей с 1 на 3 (столбцы 1 и 2) повышает качество модели. Добавление признака *m3* улучшает качество, особенно это видно для суммы платежа (прерывистая сущность) и ответчика. *Само слово* не является хорошим признаком, однако, качество сильно понизилось при замене *v3* (слово с соседями 3) на *f3* (FastText с соседями 3).

Таким образом, наименьший разброс метрики F1 для разных сущностей наблюдается на наборе *(r3,v3,m3)* с оптимизатором *pa*. Сущность *ответчик* показала наихудшее качество распознавания, по-видимому, это связано с разнообразием значений сущностей в разных документах (ли-

цо или представитель). Также видно, что качество у простых и сложных лучше, чем у прерывистых («оплата штраф/сумма»).

Table 4. Analysis of recognition quality. Refinancing agreements.

Таблица 4. Анализ качества распознавания. Договоры о рефинансировании.

Сущность	m3, r3 pa	m3, v3 pa	r1, v1, m1 pa
клиент 1	1.00	0.99	1.00
номер паспорта	0.98	0.99	0.98
номер документа	0.99	0.99	0.99
число выплат	0.98	1.00	0.98
клиент 2	0.96	0.98	0.96
сумма прощения	0.95	1.00	0.95
дата документа	0.97	0.98	0.96
день рождение	0.99	1.00	0.99
месяц оплаты	0.96	1.00	0.96
номер кредитного договора	0.98	1.00	0.98
имя банка	0.99	0.99	0.99
дата кредитного договора	0.96	0.98	0.96
итоговая сумма	0.98	1.00	0.98
дата первого платежа	0.96	0.96	0.96
день оплаты	0.98	1.00	0.98
адрес клиента	0.97	1.00	0.98
F-macro	0.98	0.99	0.98

4.2. Результаты на договорах о рефинансировании

Перейдем ко второму набору данных, в таблице 4 приведены результаты тестирования на различных наборах признаков. Обучающая выборка составила 160 документов, а отложенная — 40 документов. Стоит заметить, что метрика по всем сущностям значительно лучше, чем для предыдущего набора данных. По нашему мнению это закономерно в виду структурированности банковских документов. Как видно из таблицы, авторами был достигнут высокий результат на малом наборе данных. Однако отметим, что, поскольку документы были представлены одной организацией, то данная модель может плохо работать на документах другой.

Так как сущности в данной выборке менее вариативны, то использование самого слова как признака дает улучшение качества. Данную особенность можно увидеть во втором столбце, при использовании *v3*. Однако наиболее универсальной будет модель с признаками *(m3, r3)* (столбец 2). Последний столбец показывает, что даже используя малое количество соседей можно добиться высокого качества.

Стоит отметить, что для этих данных точного решения можно достичь с минимальным количеством трудозатрат, и поэтому он не является репрезентативным для решения задачи NER.

Заключение

В статье рассмотрен подход к решению задачи извлечения именованных сущностей из русскоязычного текста. Для решения поставленной задачи был использован метод CRF, был проведен сравнительный анализ различных алгоритмов оптимизации, алгоритм *pa* показал лучшее качество.

Авторами были использованы признаки на основе регулярных выражений, морфологии, различных векторных представлений слов, а также признаки их соседей.

Конечной целью исследования является создание универсального инструмента для извлечения различных сущностей из широкого спектра документов, поэтому в работе было проанализировано два набора данных: банковские документы имеют ярко выраженную структуру, а судебные протоколы — нет.

В ходе тестирования на различных наборах признаков было выявлено, что на неструктурированных данных качество модели растет с увеличением количества используемых соседей (но и время обучения растет полиномиально). Во время эксперимента для улучшения качества распознавания именованных сущностей хорошо себя показали признаки на основе регулярных выражений и морфологии. Заметим, что использование векторного представления существенно увеличивает время обучения и теряется возможность использования информации о соседних словах. Для структурированных документов можно получить хорошее качество даже не используя большое количество соседей.

В качестве одного из путей улучшения результатов планируется формировать дополнительные признаки. Наиболее перспективным направлением исследования является применение альтернативных подходов к извлечению именованных сущностей на основе современных нейросетевых архитектур BiLSTM и CRF.

References

[1] E. Leitner, G. Rehm, and J. Moreno-Schneider, “Fine-grained Named Entity Recognition in legal documents”, in *International Conference on Semantic Systems*, Springer, 2019, pp. 272–287.

[2] J. Straková, M. Straka, and J. Hajič, “Neural architectures for nested NER through linearization”, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5326–5331, 2019.

[3] R. Yeshpanov, Y. Khassanov, and H. A. Varol, *KazNERD: Kazakh Named Entity Recognition dataset*, 2022. arXiv: 2111.13419 [cs.CL].

[4] S. Zheng *et al.*, “Conditional Random Fields as Recurrent Neural Networks”, in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1529–1537.

[5] K. W. Church, “Word2vec”, *Natural Language Engineering*, vol. 23, no. 1, pp. 155–162, 2017.

[6] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information”, *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[7] C. Sutton, A. McCallum, *et al.*, “An introduction to Conditional Random Fields”, *Foundations and Trends® in Machine Learning*, vol. 4, no. 4, pp. 267–373, 2012.

[8] J. Lafferty, A. Mccallum, and F. Pereira, “Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data”, in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 282–289.

[9] M. Collins, “Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms”, in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Association for Computational Linguistics, 2002, pp. 1–8.

[10] S. Bird, "NLTK: The natural language toolkit", in *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, ser. COLING-ACL '06, Association for Computational Linguistics, 2006, pp. 69–72.

[11] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora", in *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*, 2010, pp. 46–50.

[12] M. Korobov, "Morphological analyzer and generator for Russian and Ukrainian languages", in *Analysis of Images, Social Networks and Texts*, Springer, 2015, pp. 320–332.

[13] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for Named Entity Recognition", *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 1, pp. 50–70, 2020.