Санкт-Петербургский государственный университет

**ПУШКАРЕВ Антон Борисович**

**Выпускная квалификационная работа**

**Восстановление полных форм числительных и сокращений для задач синтеза русского текста**

Уровень образования: магистратура

Направление 45.04.02 «Лингвистика»

Основная образовательная программа ВМ.5715. «Общая и прикладная фонетика (General and Applied Phonetics)»

Профиль «Речевые технологии»

Научный руководитель:

профессор, Кафедра фонетики и методики преподавания иностранных языков,

Скрелин Павел Анатольевич

Рецензент:

заместитель директора департамента, Общество с ограниченной ответственностью «Центр речевых технологий»,

Толстунова Екатерина Викторовна

Санкт-Петербург

2023

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ………………………………………………………….….…………4

ТЕОРЕТИЧЕСКАЯ ГЛАВА……………………………………………………...6

1. Общее устройство систем синтеза речи по тексту и значение числительных и сокращений в ходе текстовой предобработки….………6
	1. Устройство систем синтеза и принципы работы……………………...6
	2. Модуль лингвистической обработки текста и связанные с ним задачи в современных системах синтеза речи………………………………...7
		1. Проблемы нормализации текста ………………………………..8
		2. Роль числительных и сокращений в процессе нормализации текста…………………………………………………………….10
			1. Расшифровка числовых записей………………………...10
			2. Расшифровка аббревиатур и графических сокращений.11
2. Обзор существующих методов решения проблемы расшифровки нестандартных записей………………………………………………...…12

 2.1 Системы на правилах………………………………………………….13

 2.2 Подходы, основанные на статистических моделях и машинном обучении………………………………………………………………………….14

 2.3 Методы, основанные на комбинации правил и статистических моделей…………………………………………………………………………...15

 2.4 Методы на основе глубокого обучения и нейронных сетей………..17

 2.5 Выбор метода для русского языка……………………………………18

 3. Имена числительные и сокращения в русском языке………………...…20

 3.1 Имя числительное как часть речи…………………….………………20

 3.1.1 Свойства числительных и маркеры определения грамматической формы……………………………………………………….…20

 3.2 Понятие сокращений в русском языке…………………………….…24

 3.2.1 Нормативный статус графических сокращений в современном русском языке………………………………………………………………….…25

 3.2.2 Способы определения полной грамматической формы сокращений………………………………………………………………………26

ПРАКТИЧЕСКАЯ ГЛАВА……………………………………………………...27

1. Описание разработанной системы нормализации………………………27
2. Материалы исследования………………………………………………...29
3. Модуль вербализации для чисел и сокращений…………………………32

3.1 Вербализация чисел…………………………………………………...32

 3.1.1 Выведение лексемы из числа…………………………………...32

 3.1.2 Порождение грамматической формы числительных…………33

3.2 Определение верной расшифровки неоднозначных сокращений с помощью контекстных векторов…………………………………………36

1. Модуль предсказания грамматической формы …………………………39

4.1 Описание модели условных случайных полей………………………40

 4.1.1 Формальное определение CRF…………………………………40

4.2 Применение модели для предсказания формы числительных……..42

4.3 Модель CRF для определения грамматической формы сокращений44

1. Результаты ………………………………………………………………...46

5.1 Оценка работы классификатора для предсказания формы числительного……………………………………………………………..46

5.2 Оценка работы классификатора для предсказания формы сокращения………………………………………………………………..49

ЗАКЛЮЧЕНИЕ…………………………………………………….……………51

Список литературы……………………………………………………………...53

ВВЕДЕНИЕ

Настоящая работа посвящена одному из компонентов прикладных систем синтеза речи по тексту – лингвистическому модулю. Данный модуль делит входной текст на предложения, затем на слова, ставит им соответствующие лексико-грамматические метки и делает фонетическую транскрипцию с целью дальнейшего корректного прочтения и озвучивания системой. Для построения правильной транскрипции в ходе текстовой предобработки необходимо произвести нормализацию любых возможных нестандартных записей, и, в частности, работа затрагивает преобразование цифровых последовательностей и сокращений в стандартный орфографический вид. Поскольку во флективных языках числительные и сокращения могут принимать широкий спектр различных форм, цель данной работы – разработать прикладную систему, которая решит проблему выбора верной грамматической формы для русского языка, что поспособствует общей разборчивости выходной синтезированной речи.

Актуальность работы в том, что синтез речи применяется во многих сферах, в т. ч. для озвучивания чат-ботов, голосовых помощников, информации в аэропортах и вокзалах, автомобильных навигаторов, специальных приложений для людей с ограниченными возможностями и т. д. Для синтеза очень важно, чтобы текст звучал разборчиво и максимально приближенно к той форме, в какой его произносит человек. Выходной результат синтеза зависит в том числе от того, насколько «по-человечески» система озвучит те записи, которые люди предпочитают сокращать. в формат, который нужно сначала расшифровать, прежде чем произнести в устной форме.

Нормализация входного текста является сложной задачей, т.к. существует множество сокращений, цифровых последовательностей и других нестандартных с точки зрения построения транскрипции записей, которые потенциально могут иметь различные значения и грамматические формы в зависимости от контекста. В современных системах на базе русского языка данная проблема остается не до конца решенной.

Работа состоит из введения, основной части, структурированной по двум главам – теоретической и практической, заключения и списка литературы.

Теоретическая глава включает введение в проблему нормализации нестандартных записей в контексте систем синтеза речи по тексту; анализ роли числительных и сокращений в ходе текстовой предобработки; обзор работ, посвященных решению данной проблемы различными методами, а также обоснование выбора метода для русского языка; исследование русских числительных и сокращений в качестве функционирующих языковых единиц и анализ возможных маркеров определения грамматической формы.

Практическая глава включает описание материала исследования, описание предложенного метода решения задачи; описание итоговой системы нормализации числительных и сокращений, ее преимущества и недостатки; описание полученных результатов тестирования системы и анализ ошибок.

ТЕОРЕТИЧЕСКАЯ ГЛАВА

1. **Общее устройство систем синтеза речи по тексту и значение числительных и сокращений в ходе текстовой предобработки**
	1. Устройство и принципы работы систем синтеза речи по тексту

Синтез речи по тексту (TTS, от англ. Text-to-Speech) – это процесс генерации речевого сигнала из письменного текста, который является предметом активных исследований на протяжении нескольких десятилетий. Общая структура систем синтеза TTS традиционно включает модуль лингвистического анализа текста, просодический модуль и модуль генерации акустического сигнала, как описано у Dutoit. Модуль лингвистического анализа отвечает за обработку входного текста и преобразование его в фонематическое представление, которое затем обрабатывается просодическим модулем для создания соответствующих просодических характеристик, таких как интонация, ударение и ритм. Наконец, модуль генерации акустического сигнала на основе фонетических и просодических параметров, полученных на предыдущих этапах, формирует звучащую речь [Dutoit, 1997]. Существует два основных подхода к генерации сигнала: конкатенативный и параметрический. Конкатенативный синтез предполагает объединение предварительно записанных речевых единиц для формирования желаемого речевого отрезка. Параметрический синтез для генерации выходного сигнала использует математические модели на основе набора параметров [Лобанов, Цирульник, 2008]. Качество выходного сигнала системы синтеза зависит от разборчивости и естественности синтезированной речи, что, в свою очередь, зависит от эффективности работы вышеописанных компонентов системы [Lemmetty, 1999].

 В дополнение к модулю лингвистического анализа система TTS может включать отдельный модуль пре-процессинга (от англ. pre-processing) входного текста. Например, в классической книге по синтезу речи [Taylor, 2009] обсуждается важность обработки чисел и различных сокращенных записей на этапе предварительной обработки текста, поскольку эти элементы представляют трудности для систем TTS. Такая предобработка является важным шагом на этапе лингвистического анализа, так как позволяет получить более естественное и разборчивое звучание итогового речевого сигнала системы синтеза [Sproat, 1998].

1.2 Модуль лингвистической обработки текста и связанные с ним задачи в системах синтеза речи

Лингвистический модуль в системах синтеза TTS выполняет задачу преобразования входного печатного текста в его фонетическую и интонационную транскрипцию с целью дальнейшей передачи в блок акустического синтеза для генерации из нее акустического сигнала. Согласно [Рыбин, 2014], лингвистический блок обычно состоит из нескольких подмодулей, которые выполняют анализ текста на нескольких уровнях: графематическом, синтаксическом, морфологическом, семантическом. Таким образом, можно выделить основные этапы обработки текста:

* Сегментация текста на предложения;
* Токенизация (разбиение предложений на слова, слова на символы);
* Морфологический анализ (частеречная разметка, от англ. POS-tagging);
* Определение лексической и синтаксической информации для каждого слова;
* Нормализация нестандартных записей.

Важной задачей лингвистического модуля также является снятие неоднозначности во входном тексте, например, в случае омографов или слов с несколькими возможными вариантами произношения. Для этого был предложен метод автоматического разрешения омонимии, основанный на анализе контекста, в котором встречается слово [Аничкин, Чистиков, 2008]. Некоторые системы синтеза включают просодические и фонетические процессоры в лингвистический модуль. Например, в методе статистического параметрического синтеза в данный модуль входит предсказание просодии [Chu, Qian, 2001] и преобразование графем в фонемы (G2P, от англ. Grapheme-to-Phoneme) [Yao, Zweig, 2015].

Комплексные системы синтеза, основанные на нейронных сетях, позволяют в качестве входных данных напрямую использовать последовательности символов или фонем, и таким образом модуль анализа текста во многом упрощен [Tan и др., 2021]. В данном случае нормализация текста по-прежнему необходима для получения стандартного орфографического представления слова из введенных символов, а также требуется дальнейшее преобразование слов в фонемы. В некоторые комплексные модели синтеза встраиваются традиционные функции анализа текста. Например, система DeepVoice [Arık и др., 2017] реализует преобразование символов в транскрипцию на основе нейронных сетей в системе TTS, которая имитирует традиционное разделение на модули лингвистического анализа (front-end) и синтеза аудиосигнала (back-end); а некоторые работы [Stanton, Wang, Skerry-Ryan, 2018] явно предсказывают просодические характеристики с помощью текстового кодировщика.

1.2.1 Проблемы нормализации текста

Поскольку сырой печатный текст часто содержит широкий спектр нестандартных с точки зрения правил построения транскрипции записей (в зарубежной литературе называемых Non-Standart words или NSW [Sproat и др., 2001]), которые не могут быть напрямую «озвучены» системой синтеза речи, основной задачей блока предварительной обработки является нормализация, т.е. преобразование всех таких записей в стандартные слова, для которых применимы стандартные правила, используемые на этапах фонетической и просодической обработки [Zhang и др., 2019]. Как было продемонстрировано на примере флективного хорватского языка, такие записи встречаются с разной частотностью в разных типах текстов (например, больше всего – 24% в официальных документах) [Beliga, Martincic-Ipsic, 2014].

Помимо предмета данного исследования – дешифровки чисел и сокращений, Б.М. Лобанов выделяет следующие подзадачи текстового нормализатора для системы синтеза по русскому тексту:

* Очистка текста от недопустимых символов

Входной текст может содержать графические объекты, ссылки, различные маркеры, скрытые символы форматирования (кодировки) и др. неинформативные для синтеза речи символы. Для их удаления требуется база данных, содержащая только те символы, которые потенциально могут быть при необходимости нормализованы и «озвучены» синтезатором речи.

* Дешифровка специальных символов

Эта задача подразумевает правила преобразования в орфографические слова таких символов, как @, %, №, &, а также сложных математических символов «∑», «∫» и др.

* Дешифровка иностранных слов

Данная подзадача сводится к преобразованию слов, состоящих из латинских символов, таких как интернет-адреса, адреса электронной почты, названия организаций и т.д., в слова из русских букв, читаемых по общим правилам. Для этого используется база данных наиболее употребительных иностранных слов и их эквивалентов на русском языке («*Microsoft*» – «*ма́йкросо́фт*», «*www*» – «*три да́блъю*»), а также правила дешифровки латинских символов в случае прямой транслитерации.

Также в качестве проблемы исключительно русского языка выделяется корректировка буквы «ё». В большинстве случаев достаточно базы данных, содержащей все слова с буквой «ё» для поиска по ней и последующей замены. Однако, в некоторых случаях («Всё/все в машине?») для разрешения неоднозначности требуется применить семантический и прагматический анализ [Лобанов, 2011].

1.2.2 Роль числительных и сокращений в процессе нормализации текста

1.2.2.1 Расшифровка числовых записей

Декодирование числовых записей является важнейшей задачей при нормализации текста для систем синтеза речи, особенно в языках с флективным строем. Русский (наряду с другими флективными языками, такими как чешский или хорватский) имеет комплексную систему числительных [Breuillard, 1992]. Сложность связана с определением падежа, числа и рода существительных и прилагательных, которые также переносятся на зависимые от них либо согласующиеся с ними имена числительные. В русском языке разделяют два числа (ед.ч и мн.ч), 3 рода (м, ж, ср), и 6 падежей (им, род, дат, вин, тв, предл). Таким образом, поскольку порядковые числительные грамматически идентичны прилагательным, каждое порядковое числительное может иметь 36 различных морфологических форм. Как правило, числительные согласуются с родом существительного, от которого зависят. Так, "*один город*" имеет форму "*один*" в именительном / винительном падеже мужского рода, а «*одна собака*» - форму "*одна*" в женском роде в соответствии с родом существительного. Аналогично "*два города*" и "*две собаки*". В косвенном падеже, напр., предложном - числительное должно согласоваться с падежом существительного: "*в двух шагах*". Составные числительные склоняются пословно: "*с пятью тысячами пятьюстами семьюдесятью четырьмя рублями*" (тв. п.). Более подробно мы рассмотрим грамматические особенности числительных в 3 разделе данной главы.

Соответственно, наличие широкого спектра различных грамматических форм в русском языке усложняет задачу расшифровки числовых последовательностей, поскольку предполагает определение форм числительных в зависимости от контекста. Числительные также могут быть представлены в тексте в различном виде: цифры, слова, смешанные формы («*90-е*»), требующие преобразования в стандартный орфографический вид [Gorman, Sproat, 2016].

Кроме того, грамматические особенности числительных и их вариативность были предметом изучения теоретической лингвистики [Corstius, 2012; Epstein, 1978]. Знание этих особенностей и применение соответствующих формальных правил поможет в создании более надежной системы нормализации русских числительных.

1.2.2.2 Расшифровка аббревиатур и графических сокращений

Сокращенные формы записи часто встречаются в печатных текстах. В то время как человек легко может извлечь и восстановить полную форму по смыслу или контексту, автоматические системы синтеза речи нуждаются в алгоритмах извлечения полной формы сокращения. Дополнительно сокращения могут иметь несколько значений в зависимости от контекста, что делает их декодирование сложной задачей [Ng, Gorman, Sproat, 2017].

Сокращения в большом количестве представлены в газетных и новостных текстах, научно-популярной и в меньшей степени в художественной литературе [Beliga, Martincic-Ipsic, 2014]. Включают различные комбинации букв, цифр, специальных небуквенных символов: «*км, и.о., мск, Гб, м/с, м2, C#*», а также специальные знаки: *«%, °, $, №».* Важно учесть наличие точки, чтобы верно построить алгоритм деления текста на предложения. Расшифровываются за счет анализа соседних слов. Флективные языки, такие как русский, создают дополнительные трудности при расшифровке. Необходимо учесть семантическую неоднозначность: в зависимости от контекста «*м.*» может обозначать „*метр*“ или „*метро*“; «*ст.*» — „*станция*“ или „*статья*“, или совпадать с несокращенными словами, например, «*Кб*» — „*килобайт*“ или аббревиатурой КБ („*конструкторское бюро*“), «*им.*» — „*имени*“ или личным местоимением.

Для снятия подобной неоднозначности осуществляется поиск слова или другого элемента, ключевого для расшифровки: 2012 г. („*год*“), г. Псков („*город*“), ст. 105 УК РФ („*статья*“), ст. Москва-Сортировочная („*станция*“) и т.п. [Хомицевич, Рыбин, Аничкин, 2013]; основное значение в таких случаях при выборе верной формы имеет наличие числительного слева (1 км „*километр*“, 2 км „*километрА*“, 12 км „*километрОВ*“) и наличие предлога слева, в том числе перед числительным (более 1 км „*километра*“, к 1 км „*километру*“, до ст. Бологое „*станции*“).

[Żelasko, 2017] исследовали, достаточно ли морфосинтаксических тегов для расшифровки сокращений в языках с выраженной флективностью. Исследование показало, что хотя морфосинтаксические теги могут помочь, для точной расшифровки в таких языках требуется дополнительная контекстуальная информация.

В исследовании [Zelinka, Romportl, Müller, 2011] авторы предложили статистический метод расшифровки сокращений, который использует частотные распределения и контекстную информацию для определения наиболее подходящей полной формы для данного сокращения, что значительно повысило точность расшифровки сокращений при нормализации текста.

**2. Обзор существующих методов решения проблемы расшифровки нестандартных записей**

Ранние работы по нормализации текста основаны на системе на правилах [Allen и др., 1987], в то время как последние работы используют нейронные сети для моделирования нормализации текста в виде «от последовательности к последовательности», где в качестве исходной последовательности выступает нестандартная запись, а в качестве целевой – нормализованная вручную запись [Mansfield и др., 2019; Sproat, Jaitly, 2016]. В последнее время в некоторых работах [Zhang и др., 2020] предлагается объединить преимущества моделей, основанных на правилах и нейронных сетях, для дальнейшего повышения эффективности нормализации текста. Далее рассмотрим каждый подход подробнее.

2.1 Системы на правилах

До XXI в. большая часть работы по нормализации текста в большинстве систем TTS выполняется с использованием созданных вручную наборов правил, адаптированных к конкретным областям применения [Allen и др., 1987; Black, Taylor, Caley, 1999]. Например, в различных приложениях системы синтеза AT&T Bell Labs было сочтено важным уметь правильно определять и произносить (американские и канадские) телефонные номера [Sproat, 1998]. Такой детектор телефонных номеров был включен как часть системы предварительной обработки текста. Однако, такие случаи можно было назвать лишь частичным покрытием проблемы, что считалось нормой для систем синтеза речи по тексту. Неоднозначные расшифровки — например, “*St*” как *saint* или *street* (*st.*) — обычно обрабатывались правилами, учитывающими особенности контекста. В данном конкретном случае, если следующее слово начинается с заглавной буквы, то вполне вероятно, что правильное прочтение - *Saint* (*St. John*), если же предыдущее слово начинается с заглавной буквы, то верное прочтение *Street (Fleet st.)*.

Основной недостаток данного подхода - необходимо первичное представление о том, какие конкретные случаи должны обрабатываться, что часто нецелесообразно, особенно при переходе на другой жанр текста, поскольку даже в хорошо изученных областях, таких как тексты новостных лент, часто встречаются новые примеры нестандартных записей [Sproat и др., 2001]. Еще одна сложность касается морфологически богатых языков, где в зависимости от свойств соседних слов из контекста нестандартные записи могут быть по-разному расшифрованы. Так, для русского языка был применен морфологический и синтаксический анализ непосредственного контекстного окружения слова (цифры, знака) в работе [Лобанов, 2011] и более поздней работе [Хомицевич, Рыбин, Аничкин, 2013] с целью выбора верного варианта прочтения сложных элементов, в т.ч., цифровых записей и сокращений. Данный алгоритм построен на правилах, которые эксперт-лингвист может настроить в отдельных файлах. Эксперимент показал, что такой метод позволяет корректно разрешить подавляющее большинство случаев неоднозначности прочтения. В качестве материала использовалась выборка текстов новостного (публицистического) типа. В результате правильно были прочитаны 95% сложных элементов. Однако, авторы установили, что основным источником ошибок являются сложные случаи для анализа:

 1) в случае недостатка информации в ближайшем контексте предполагается анализ более отдаленных синтаксических связей. К примеру, во фрагменте: „*выбирать между 154 млрд кубометров по более низкой цене и 150 млрд по более высокой*“ второе числительное отделено несколькими другими членами предложения от относящегося к нему предлога;

2) ошибочные или нестандартные формы записи, например, «*в 300-стах метрах*» вместо «*в 300-х метрах*»; «*437 доллара*» вместо «*437 долларов*»; 3) формы записи, изначально не предназначенные для чтения вслух, такие как многосложные цифровые записи, слова, полностью или частично замененные звездочками и т.п.

2.2 Подходы, основанные на статистических моделях и машинном обучении

Среди статистических методов, примененных к русскому языку для задачи нормализации цифровой записи числительных, - аппарат условных случайных полей [Романенко, 2014]. В качестве обучающего материала было отобрано 8251 предложение из Национального корпуса русского языка, а эксперимент проводился на 2017 тестовых предложениях, содержащих цифровые записи. За основу методики была взята модель определения переменной как числительного по следующим признакам: GRAM – грамматические метки слов; SPEL - метки особенностей написания числительного (длина в символах, последняя цифра числительного); SPEC – является ли слово «характерным» для употребления с количественными или порядковыми числительными; NEAR – признаки соседних слов. Далее по данным признакам выделяется числительное, которое дополняется грамматическим описанием из 5 следующих меток: TYPE – тип числительного; CASE – падеж; GEND – род; SNGL – число; ANIM – одушевленность числительного. Метрика оценки полученного результата Acc рассчитывалась как число верных ответов, деленное на общее число ответов, и данный показатель составил 92,39% точности определения грамматических характеристик, усредненно по всем меткам.

Практически одновременно была реализована система нормализации русского текста Normatex при помощи контекстно-зависимых преобразователей с конечным числом состояний. Покрыты количественные и порядковые числительные, аббревиатуры, графические сокращения, корректно расшифровываются телефонные номера, даты, коды, и т.д. [Lukanin, 2015]. Данная система оценивалась на разработанном параллельном тестовом корпусе, в которым все цифры (2511 цифр = 977 числовых токенов), сокращения (379) и аббревиатуры (431) вручную расшифрованы в полные слова или заменены произносимыми буквосочетаниями (например, инициалы заменены транскрипцией букв русского алфавита). В итоге были корректно расшифрованы 920 из 970 чисел (94,17%). Среди ошибок автор выделяет перечисления, нешаблонно прописанные телефонные номера (“*8-35146)-30383”*), даты (“*2012.11.07”*), временные интервалы (“*Oбeд 14.00–15.00”*), а также номера документов (“*Инcтpyкция № 3-1-0 г.*”). Самая высокая точность (98,05%) была получена при расшифровке сокращений. Таким образом, общая точность составила 93,95%.

2.3 Методы, основанные на комбинации правил и статистических моделей

В исследовании, посвященном нормализации русских имен числительных, R. Sproat представил метод частично контролируемого обучения. В работе описывается комбинация методов, основанных на правилах и обучающих данных, для обучения систем синтеза речи декодированию числовых последовательностей на русском языке [Sproat, 2010]. Автор объяснил низкий интерес к данной проблеме отсутствием больших объемов аннотированных данных, необходимых для обучения языковых моделей. Поэтому в качестве материала авторами была создана база данных из нескольких миллионов расшифрованных числительных и сопоставлена со строками, содержащими цифры. Были собраны все экземпляры последовательностей одного или нескольких слов из списка числительных на всех русскоязычных страницах из некоего «среза Сети Интернет». Также было собрано окно из трех токенов по обе стороны от числительного (контекст слева и справа). Общий алгоритм обработки можно описать следующим образом:

* предоставить начальный список L всех верных форм односложных числительных;
* искать на веб-страницах последовательности терминов из L вместе с их контекстами;
* используя контекстно-свободную грамматику числительных, отфильтровать полученный список комбинаций, соответствующий общим ожидаемым свойствам корректно сформированных числительных: грамматика реализована как преобразователь с конечным числом состояний (от англ. Finite State Transducers, FST), который будет принимать только обоснованно выглядящие вербализации чисел и сопоставлять их с соответствующими последовательностями цифр;
* результатом предыдущего шага является большой список аннотированных пар «строка цифр – числительное» в контексте. Затем эти данные используются для обучения модели порождать контекстно подходящее числительное исходя из данной строки цифр.

Авторы сравнили генеративные языковые модели, основанные на n-граммах, и два различительных дискриминационных метода (перцептрон и список решений) и продемонстрировали, что n-грамм модель дает наилучшую производительность (процент ошибок 0.12). Кроме того, было обнаружено, что почти все ошибки в изученной системе связаны с проблемами выбора правильной контекстной расшифровки, в то время как сравнительно малое количество ошибок (0.08) связано с изначально неверно подобранной основой числительного.

2.4 Методы на основе глубокого обучения и нейронных сетей

Среди методов машинного обучения, применяемых в системах нормализации нестандартных записей, стоит особо выделить методы глубокого обучения, стремительно набирающие популярность в последние годы из-за возросших вычислительных мощностей компьютеров, а также существенно выросших объемов данных, которые можно использовать в качестве обучающего материала, – рекуррентные нейронные сети [Sproat, Jaitly, 2016], сверточные нейронные сети [Yolchuyeva, Németh, Gyires-Tóth, 2018].

Согласно наблюдениям исследователей, задачу нормализации текста в контексте глубокого машинного обучения целесообразно разбить на три основные подзадачи:

1) разбиение текста на сегменты (токены), каждый из которых является:

а) обычным словом, не требующим нормализации;

б) пунктуационным знаком, который удаляется;

в) экземпляром одного из форматов (семиотических классов) нестандартных записей (время, дата, адрес, количественное числительное, аббревиатура и т.д.);

2) определение набора возможных расшифровок (вербализаций) для входного токена типа в) ;

3) вычисление наиболее контекстно подходящей из них [Zhang и др., 2019].

Если более ранние исследования были сосредоточены на нормализации на уровне слова (или фразы), то архитектура нейронных сетей позволяет напрямую нормализовать целые предложения, используя в качестве входных и выходных данных последовательность символов [Mansfield и др., 2019]. В связи с этим к задаче нормализации применялись методы из другой области обработки естественного языка – машинного перевода, где исходная и целевая запись представляют собой печатный и нормализованный текст соответственно. Результирующий показатель ошибок WER (от англ. Word Error Rate) – 0,17% [Mansfield и др., 2019; Yolchuyeva, Németh, Gyires-Tóth, 2018].

[Sproat, Jaitly, 2016] исследовали две нейронные модели нормализации текста на материале русского и английского языков. Первая включает в себя архитектуру рекуррентной нейронной сети LSTM, которая создает список-граф возможных расшифровок, которые затем пересчитываются отдельной языковой моделью LSTM. Вторая модель — сеть RNN с механизмом внимания, работает по принципу «от последовательности-к-последовательности», успешно применялась для распознавания речи [Chan и др., 2016]. Последующее исследование сосредоточено исключительно на последней модели, которая превзошла первую [Ng, Gorman, Sproat, 2017]. Тем не менее, она по-прежнему допускает значительное количество грубых ошибок, критических для подобного уровня систем. Если система «прочтет» словосочетание «*45км*» как «*тридцать пять километров*», фактическая информация сообщения исказится. Sproat и др. обнаружили, что нейросетевые модели особенно подвержены такого рода ошибкам, и назвали их «непоправимыми». Отмечается, что даже если непоправимые ошибки случаются редко, они носят спорадический характер и являются существенным недостатком данного подхода к нормализации [Zhang и др., 2019].

2.5 Выбор метода для русского языка

Как показали эксперименты с нормализацией текста на основе глубоких нейронных сетей, несмотря на в целом довольно высокую точность, значительная часть ошибок с цифровыми последовательностями, в т.ч. на примере русского языка, связана с тем, что система верно определяет лемму числительного, но в неподходящей грамматической форме. Для исправления подобного рода ошибок вводится дополнительный слой корректирующих грамматик (от англ. Covering Grammars). В целом такие ошибки менее опасны, чем неверное определение основы: если система ошибется только в определении грамматической формы слова, носитель языка все равно сможет восстановить предполагаемое значение [Sproat, Jaitly, 2016]. Однако, вышеупомянутых «непоправимых» ошибок определения начальной формы не удается избежать полностью, что является критичным для системы нормализации, в отличие от, например, системы машинного перевода. В связи с этим целесообразно для определения начальной формы русских числительных использовать систему, основанную на правилах, которая исключит подобные грубые ошибки.

Как показал обзор литературы, статистический подход, основанный на условных случайных полях (CRF), не требует больших объемов данных и значительных вычислительных мощностей относительно методов глубокого обучения, и при этом успешно применялся на материале русского языка для определения грамматической формы числительных [Muzychka, Romanenko, Piontkovskaja, 2014], а также для снятия морфологической омонимии с помощью контекста в качестве части нейронной модели [Shen и др., 2016]. В обоих случаях итоговая точность схожа (91.06% против 91.13%). Таким образом, использование комбинации системы на правилах и статистического подхода на основе CRF-модели потенциально позволит избежать ошибок определения основы числительного, а также, вероятно, приблизит итоговую точность определения конечной грамматической формы к показателям нейронных сетей, при этом используя значительно меньше обучающего материала и вычислительных мощностей.

Также есть основания полагать, что данный метод можно успешно применить и для русских сокращений, т.к. он продемонстрировал высокую точность расшифровки сокращений на примере другого флективного языка – чешского [Zelinka, Romportl, Müller, 2011].

**3. Имена числительные и сокращения в русском языке**

3.1 Имя числительное как часть речи

Согласно русской грамматике, имя числительное – это часть речи, обозначающая количество и выражающая это значение в морфологических категориях падежа и рода [Брызгунова и др., 2005]. С точки зрения данной работы нас интересуют числительные, которые на практике в печатном тексте чаще всего записываются в виде цифровых последовательностей. Поэтому целесообразно опустить описание собирательных *(«оба»)* и неопределенно-количественных *(«много»)* числительных. Термин «числительное» традиционно употребляется также для наименования всех слов с количественно-числовым и счетно-порядковым значениями. Такие слова отвечают на вопросы «сколько?» и «который?»: *два, три, тысяча, миллион, нуль (и ноль); первый, второй, десятый, сотый*; принадлежат к разным частям речи: к числительным, существительным, прилагательным; их объединяет то, что все они являются *счетными словами* и могут быть записаны в сокращенном виде как последовательность арабских либо римских цифр, а также в виде графического сокращения: *тыс., млн., млрд.*, поэтому нуждаются в нормализации.

3.1.1 Свойства числительных и маркеры определения грамматической формы

Все счетные слова обладают некоторыми общими грамматическими особенностями: числительные не имеют противопоставления по числу внутри парадигмы и, следовательно, не имеют словоизменительной грамматической категории числа. Например, счетно-местоименное слово *один* (в количественном значении) отличается от прилагательных тем, что может соединяться только с существительными в ед. числе, за исключением наименований парных предметов *(«одни сани, одни сутки»)*, таким образом имея внешние признаки мн. ч. [Зализняк, 2022] Также «*один»* участвует в формировании составных числительных (*одна* *тысяча двести тридцать один*), в которых (во всех формах ед. ч.) совмещает в себе формальные признаки прилагательного с функциями количественного числительного.

Таким образом, среди всех слов со счетно-количественным значением присутствует сильная грамматическая неоднородность: различия в морфологических категориях и парадигматике таких слов, как «*два»* и «*второй*» определяют их принадлежность к разным частям речи. Счетно-порядковые прилагательные типа *«второй, третий, десятый*» (называемые иногда также порядковыми числительными) не имеют своих собственных, отличающих их от прилагательных морфологических признаков. С другой стороны, такие слова, как «*миллион, миллиард, нуль»,* не имеют морфологических признаков, отличающих их от существительных (т.е. имеют формы обоих чисел). Счетно-порядковые прилагательные и счетно-местоименное слово *«один»* во всех падежах с определяемым существительным последовательно согласуются [Брызгунова и др., 2005].

Не имея категории числа, числительные избирательно относятся к формам числа тех существительных, с которыми сочетаются. В зависимости от сочетаемости с формами ед. или мн. ч. существительных числительные делятся на две группы: 1) числ. «*два, полтора, три, четыре*», которые соединяются с сущ. в ед. ч. (род. п.): *две тысячи, три миллиона;* 2) числ. *пять, шесть, семь* и далее, которые соединяются с сущ. во мн. ч. (род. п.): *пять миллиардов*. Правило действует для им. и вин. падежей; во всех других падежах сочетающееся с числительным существительное последовательно употребляется в форме мн. ч.: *два триллиона*, но *двух триллионов*.

К составным числительным причисляются т.н. дробные числительные типа *две пятых* (подразумеваются доли единицы), *семь двадцатых, девять тридцать вторых; одна целая шесть десятых, три целых и двадцать пять сотых*, которые представляют собой имеющие количественное значение сочетания слов, относящихся к разным частям речи. Существительное при таком сочетании слов всегда имеет форму род. п. ед. ч. (слова pluralia tantum — род. п. мн. ч.): *«пять десятых участка», «три пятых суток»; «пяти десятым участка».* Данные характеристики существительного можно использовать как маркер-определитель категории сочетания слов с дробным значением.

Непосредственно числительные, кроме слов «*два*» и «*полтора*», не имеют морфологической категории рода. Для данных исключений формы муж./ср. или жен. рода присутствуют только в им. и вин. падежах и определяют существительные соответствующего рода *(«два стола», «полторы булки»*). В остальных падежах имеют одинаковую форму. Остальные числительные никак не указывают на грамматический род тех существительных, с которыми они сочетаются: «*четыре стола», «четыре книги», «четыре окна»* [Брызгунова и др., 2005]*.* Таким образом, род указанных исключений можно определить по роду согласованного существительного.

Категория падежа у числительных — это словоизменительная морфологическая категория, представленная шестью рядами падежных форм, по флексиям совпадающих с падежными формами существительных или прилагательных. Склонение числительных по падежам в академической грамматике представлено несколькими типами:

1. числ. «*два», «три», «четыре*»: имеют только флексии мн. ч. Образуют падежные формы по образцу смешанного склонения прилагательных.
2. от «*пять*» до *«десять*», а также числ. на «-*дцать*» и «-*десят*»: имеют только флексии ед. ч. Образуют падежные формы по 3-му склонению существительных. В числительных на *«-десят*» склоняются обе составные части («*пятидесяти», «пятьюдесятью*»)
3. числ. «*двести», «триста», «четыреста*» и все числ. на «*-сот*»: имеют только флексии мн. ч. Образуют падежные формы так же, как сущ. 1-го скл., мн. ч. У всех числ. на *«-сот*», так же как у числ. на *«-десят*», склоняются обе части сложения. Вторая часть основы числ. «*двести*», *«триста»* и «*четыреста*» в форме род. п. имеет фонемный состав */сот/*, в остальных падежах — */ст/*; у числ. на *«-сот»* («*пятьсот, девятьсот»*) в формах дат., тв. и предл. падежей вторая часть основы имеет фонемный состав */ст/*;
4. числ. «*сорок», «девяносто», «сто», «полтора*»: имеют только флексии ед. ч., а также только две различающиеся падежные формы:

а) форму им. и вин. п. (для «*сорок*» - нулевая флексия; для «*девяносто*» и «*сто*» - флексия -|о|);

б) форму, единую для всех остальных косв. падежей (для «*сорок*» - флексия -|а|; для «*девяносто*» и «*сто*» - флексия -|а|).

Отдельно стоит отметить составные числительные, в которых при образовании падежных форм в письменной речи должны склоняться все слова, входящие в состав, по вышеуказанным правилам (Русская грамматика, Брызгунова, 2005).

Числительные непоследовательно обозначают одушевленность/неодушевленность тех существительных, с которыми они сочетаются. Из количественных числительных на одушевленность указывают только слова «*два, три и четыре*»: «*видел две (три) картины*», но «*двух товарищей*», «*поднял четыре кубика*», но «*поднял четырех котят*». Таким образом, категорию одушевленности данных числительных можно определить по этой же категории согласованных существительных.

Среди синтаксических связей, влияющих на категорию падежа числительных, отмечается, что все они в им.п. управляют род.п. без предлога: «*десять домов», «пять братьев*». При словах «*два(две)*», «*три*», «*четыре*», а также составных числительных с данными последними компонентами в форме им.п. - имя в род.п. стоит в форме ед.ч.: «*два миллиона», «две тысячи*» [Брызгунова и др., 2005]. В сочетаниях с предлогами при изменении падежной формы числительное согласуется с падежной формой имени («*без двух столов»*). Соответственно, в данной структуре «*предлог-числ.-сущ*» падеж неизвестного числительного можно определить по падежу согласованного имени. Исключением является сочетание с предлогом «*по*», когда числительное выступает в форме, совпадающей с формой им.п., и управляет существительным, а не согласуется с ним («*дали по два карандаша», «по двое ножниц», «по девяносто мест», «по триста рублей*»).

В формах род., дат., тв., и предл. падежей существительные согласуются с количественными числительными в падеже: *«на семи ветрах», «со ста пятьюдесятью девятью солдатами».*

Количественное числительное может одновременно управлять двумя род. падежами, при этом первая из управляемых форм, называющая единицу исчисления, оказывается информативно излишней (избыточной): она приобретает указательную (местоименную) функцию: «*десять штук патронов», «пять человек солдат»*. Эта структура может быть полезна при определении падежа сокращения, если одно из зависимых имен сокращено: «*10 шт. патронов*».

Также отмечается, что порядковые (счетные) прилагательные могут иметь при себе зависимые падежные формы: «*первый из пришедших*», «*третий от начала*»; и наречия: «*второй справа*», «*третий подряд*» [Брызгунова и др., 2005].

3.2 Понятие сокращений в русском языке

Согласно В.Ф. Ивановой, «графические сокращения являются специфическим актом письма. Они не являются самостоятельными словами; графические сокращения не имеют своего фонемного состава: при чтении графическое сокращение заменяется тем словом, сокращением которого оно является: Спб. читается как Санкт-Петербург, г. – как год. Но расшифровываются графические сокращения лишь эпизодически» [Иванова, 1976].

Сокращения бывают также неграфическими, выступающими в качестве полноценных единиц устной и письменной речи и имеющими свой фонемный состав. К ним относятся аббревиатуры: «РФ», «вуз»; сложносокращенные слова: «филфак», «колхоз», «зарплата». Однако, в рамках данной работы мы будем рассматривать только те сокращения, которые не имеют собственных правил чтения в устной речи и потому нуждаются в автоматической расшифровке в рамках системы синтеза речи по тексту. Такие сокращения используются преимущественно в письменной (печатной) речи из потребности в скорописании и с целью экономии времени и усилий. Работа также не затрагивает т.н. «авторские» сокращения, образованные пишущим «на ходу».

3.2.1 Нормативный статус графических сокращений в современном русском языке

В зависимости от графических приемов образования, В.Ф. Иванова выделяет следующие типы сокращений:

1. точечные: «г.» - год, «т.» - тонны, «р.» - рубли;
2. дефисные: «кол-во» как количество, «р-н» - район;
3. косолинейные: «км/ч» - километров в час, «н/д» - нет данных, нет доступа, «б/у» - бывший в употреблении, «к/ф» - кинофильм;
4. курсивные, выделяются особым шрифтом – курсивом: *г* – грамм, *л* – литр;
5. нулевые сокращения, графически никак не выделяются на письме: «с» - секунда, «м» - метр, «кг» - килограмм;
6. комбинированные: «ж.-д.» - железнодорожный, «об/мин.» - оборотов в минуту [Иванова, 1976].

При составлении словаря сокращений в качестве части прикладной системы нормализации в данной работе мы будем опираться на общепринятые стандарты сокращений слов на русском языке в библиографической записи, закрепленные в ГОСТ 7.0.12-2011 [ГОСТ, 2011]. Однако, в обучающих данных некоторые типы сокращений могут выделяться по-другому, либо не выделяться согласно стандартам, например, курсивные сокращения представляются как нулевые; нулевые – как точечные и наоборот. В таких случаях система попытается предсказать то сокращение, которое подразумевалось исходя из контекста.

3.2.2 Возможные способы определения полной грамматической формы сокращений

Как отмечает В.Ф. Иванова: «Для того, чтобы сокращение было понято читающим, пишущий должен учитывать следующие факторы:

1. какова минимальная часть слова, которая может передать читающему смысл слова в целом;
2. насколько вероятно по условиям контекста появление того слова, которое сокращает пишущий. Чем вероятнее появление того слова, которое сокращает пишущий, тем в большей степени может быть произведено сокращение (большая вероятность появления слова увеличивает возможность догадки читающего);
3. графические принципы сокращения слов»[Иванова, 1976].

Правила оформления русских графических сокращений впервые были сформулированы в статье «О русских графических сокращениях» [Санников, 1964]. Среди основных принципов отмечается:

* при сокращении нельзя опускать начальную часть слова, следует опускать минимум две буквы (искл. «ю.» - юг по аналогии «с.» – север);
* сокращенная часть не может оканчиваться на гласную и на «й», «ъ», «ь»;
* допускается опущение линейно не расчлененной последовательности букв – символов, следующих друг за другом (неправ. «фбрка» - фабрика; исключения: «млн» - миллион, «млрд» - миллиард, а также дни недели – «пн.» - понедельник).

ПРАКТИЧЕСКАЯ ГЛАВА

**1. Описание разработанной системы нормализации**

В данной работе предлагается система нормализации чисел и сокращений в тексте, реализованная на языке программирования Python версии 3.9. Алгоритм нормализации текста представляет собой следующую последовательность действий:

1. программа принимает на вход текст, содержащий ненормализованные числительные и/или сокращения;
2. текст токенизируется;
3. для каждого целевого числительного и/или сокращения определяется его грамматическая форма — теги, описывающие значения грамматических категорий;
4. для каждого целевого числительного и/или сокращения определяется его лемма;
5. для каждого целевого числительного и/или сокращения производится преобразование леммы, полученной на предыдущем этапе, в словоформу, соответствующую форме, предсказанной в 3 пункте.
6. каждое нормализованное числительное и сокращение подставляется в выходное предложение на изначальное место. Нормализованный текст подается на выход программы.

Для реализации данного алгоритма были поставлены следующие задачи:

1. подготовить выборку данных для обучения статистической модели предсказания грамматической формы числительного/сокращения;
2. разработать систему, основанную на правилах, которая будет соотносить цифровые записи числительных с начальной формой числительного (леммой) и подставлять ее на выход. Например, "*45*" = "*сорок пять*";
3. разработать систему соотношения сокращенного варианта записи слова и его полного варианта на основе словаря сокращений:
	1. для однозначных сокращений;
	2. для неоднозначных сокращений. Такая система основана на алгоритме снятия неоднозначности на основе векторных представлений контекстов и лексической базы данных RuWordNet;
4. обучить модель условных случайных полей (Conditional Random Fields, CRF) для определения грамматической формы числительного. На вход модели подается текст, содержащий ненормализованные числительные, выход модели представляется в виде промаркированной последовательности, в которой каждому числительному приписан тег, соответствующий его грамматической форме. Все остальные слова маркируются специальным тегом ‘O’;
5. Обучить модель условных случайных полей (CRF) определять грамматическую форму сокращения. На вход модели подается текст, содержащий ненормализованные сокращения, выход модели представляется в виде промаркированной последовательности, в которой каждому сокращению приписан тег, соответствующий его грамматической форме. Все остальные слова маркируются специальным тегом ‘O’;
6. разработать модуль, который будет проводить финальную нормализацию текста — подставлять в целевой текст нормализованные числительные и сокращения, представленные в виде словоформы, полученной в результате преобразования леммы в грамматическую форму, предсказанную моделью CRF. Пример работы данного модуля показан на рисунке 1. Добавить специальное окно-таблицу, куда будут выводиться «обнаруженные системой» числа и сокращения, их основы и предсказанные метки с целью анализа ошибок.



**Рис. 1 – пример работы итоговой системы нормализации текста**

1. Протестировать результат работы системы и оценить эффективность определения грамматических характеристик числительных и сокращений; проанализировать ошибки.

**2. Материалы исследования**

В работе использовались материалы национального корпуса русского языка (НКРЯ)[Национальный корпус русского языка. 2003-2023], созданного институтом русского языка им. В. В. Виноградова РАН (ИРЯ РАН). Корпус содержит прозаические письменные тексты на современном русском литературном языке, взятые за период с XVIII в. до нашего времени общим объемом 374 млн слов. Коллекция материалов сбалансирована по жанрам и состоит из художественных, научных, публицистических текстов, а также расшифровок устной речи и материалов Интернета. Аннотация корпуса представлена на нескольких уровнях: морфологическом, лексическом и мета-текстовом.

Для обучения и тестирования прикладной системы использовалась часть корпуса, доступная по лицензионному соглашению, со снятой вручную морфологической омонимией и дополнительной коррекцией результатов автоматического морфологического анализа. Подкорпус включает тексты за период с 1950г. – начало XXIв. объемом 1 млн словоформ (~500текстов). Для эксперимента были отобраны контексты «публицистика», «реклама», «электронная коммуникация», «наука», где цифровые записи и сокращения представлены в наибольшем объеме [Beliga, Martincic-Ipsic, 2014].

 Каждая словоформа имеет проверенный вручную набор морфологических меток, включая указание лексемы и грамматические характеристики. В ряде случаев представлены альтернативные интерпретации словоформ. Целевые в контексте исследования нестандартные записи – числа и сокращения – помечены тегами <*ciph*> и <*abbr*> соответственно. Однозначное указание леммы сокращенной записи в данном контексте (например, «*т. - тонн*», «*т. – то*», «*т. – так*» или «*т. – телефон*») важно при составлении словаря сокращений, а также создании и расширении обучающей выборки для разрабатываемой модели. Пример аннотации корпуса в формате *xml* приведён на рисунке 2.

 

***Рис. 2 – Фрагмент аннотации корпуса НКРЯ***

При составлении морфологической разметки (часть речи, род, падеж и т.д.) создатели корпуса опирались на морфологическую модель, представленную в «Грамматическом словаре русского языка» [Зализняк, 2003].

Дополнительно в работе использовался корпус OpenCorpora, представляющий собой открытую коллекцию текстов на русском языке с морфологической разметкой [Открытый корпус русских текстов, 2023]. Общий объем корпуса составляет ~2млн словоформ. Важным аспектом для исследования числительных и сокращений является наличие подкорпуса с ручной дезамбигуацией (разрешением) морфологической омонимии в количестве ~100тыс. словоформ. В корпусе представлены разнообразные тексты, включая художественную литературу, новости, блоги и научные статьи, что позволило оценить эффективность обработки чисел и сокращений в разных контекстах. OpenCorpora содержит информацию о числительных, представленных в орфографическом виде; сокращениях и их морфологических свойствах, таких как часть речи, падеж, род, число и др., что поспособствовало расширению обучающей выборки для разрабатываемой модели и обеспечило более эффективное обучение модели предсказанию грамматической формы. Пример аннотации корпуса приведен на рисунке 3.

***Рис. 3 – Аннотация корпуса OpenCorpora***

**3. Модуль вербализации для чисел и сокращений**

3.1 Вербализация чисел

Вербализация чисел предъявляет высокие стандарты к точности, поскольку, если системе синтеза речи можно «простить» некорректную расшифровку многозначного сокращения (напр. “*plz*” как “*plaza*” вместо “*please*”), то в случае с числами расшифровка «*35*» как «*семидесяти девяти*» полностью меняет смысл предложения и крайне нежелательна. По этой причине в настоящей прикладной системе было решено вручную формализовать правила замены чисел на русские числительные.

3.1.1 Выведение лексемы из числа

С целью реализации выведения начальной формы числительного в составе системы нормализации был создан отдельный модуль *“NumberLemma\_conversion”* на языке Python, который принимает на вход число от 1 до 15 цифр и осуществляет его перевод в текстовую форму. Для этого входное число разбивается на отдельные цифры или группы цифр в зависимости от их количества, и далее с помощью отдельных методов осуществляется последовательный перевод каждого разряда числа в соответствующие числительные.

Например, методы *Lemma0\_9* и *Lemma10\_19* заменяют однозначные и двузначные цифры на простые числительные «восемь», «двенадцать» и т.д; методы *Lemma20-90* и *Lemma100-900* заменяют разряды десятков и сотен на простые («*сорок», «сто*») и сложные числительные («*семьдесят», «пятьсот*»). Методы *Lemma\_6\_Digits*, *Lemma\_9\_Digits*, *Lemma\_12\_Digits* и *Lemma\_15\_Digits* обрабатывают классы тысяч, миллионов, миллиардов и триллионов, разбивая входное число на группы по 3 цифры соответственно каждому классу, передавая их в вышеуказанные методы и добавляя существительные со значением числа «*тысяча*», «*миллион*», «*миллиард*», «*триллион*» - таким образом формируя составные числительные. В случае если ввод содержит более 15 цифр, модуль интерпретирует его как последовательность цифр и переводит каждую цифру в простое числительное, используя регулярные выражения (в качестве части нормализации текста описаны в работе [Jurafsky, Martin, 2019]).

Модуль поддерживает обработку чисел с плавающей точкой/запятой, заменяя разделитель на слово «*целых*», формируя из дробной части числительное по алгоритму выше, а также прибавляя к нему соответствующую десятичную приставку «*десятых*», «*сотых*» и т.д., вплоть до 15 цифр после разделителя.

Модуль также формирует числительные из числителя и знаменателя обыкновенных дробей (записанных через «/»), где знаменатель представляет собой порядковое числительное. Перевод чисел в порядковые числительные реализован с использованием существующего python модуля “*num2words*” версии 0.5.12.

3.1.2 Порождение грамматической формы числительных

Для порождения словоформы числительного разработан модуль на языке python, который получает сформированную на предыдущем этапе лемму числительного и порождает необходимую грамматическую форму с использованием морфологического анализатора для русского языка*“pymorphy2”* версии 0.9.1.

Для демонстрации данного этапа был разработан графический интерфейс с помощью библиотеки “*tkinter*”, где пользователь может ввести любое число с учетом ограничений из пп.3.1.1, задать для него следующие морфологические категории:

* Тип (количественное или порядковое);
* Падеж
* Число (при наличии)
* Род (при наличии)

и получить числительное в требуемой форме на выходе. Последние 2 категории у количественных числительных присутствуют только для счетно-местоименного слова «*один*»; у числительного «*два*» добавляется только категория рода. Пример работы модуля представлен на рисунке 4.



**Рис. 4 – демонстрация порождения формы числительных**

Система учитывает основные правила склонения числительных:

* в составных числительных изменяется по падежам каждое слово;
* счетно-порядковые прилагательные (порядковые числительные) склоняются как прилагательные;
* в составе порядковых числительных склоняется только последнее слово

[Брызгунова и др., 2005] - с некоторыми допущениями. Например, счетное существительное «*тысяча*» без определения в Тв.п. имеет форму «*тысячью*», однако настоящая система всегда подставляет определение («*одной тысячей*»).

В особом порядке прописаны правила склонения существительных с счетным значением «*миллион, тысяча*» и т.д. путем введения для них отдельно категории числа в зависимости от предшествующего числительного (единственное, если это «*один, два, три, четыре*»; также родительный падеж после числительных «*два, три, четыре*» в случае, если общий выбранный падеж – именительный или винительный: «*три миллионА*»; во всех остальных случаях присваивается множественное число и падеж, выбранный пользователем).

Реализовано склонение дробных числительных, полученных из чисел с плавающей точкой/запятой, для которых программа корректно определяет форму прилагательного «*целый*», а также числительных-прилагательных «*десятый*», «*сотый*», «*тысячный*» и т.д. в зависимости от предшествующего числительного. Если это «*один*», «*два*», «*три*», «*четыре*», то в именительном и винительном падежах десятичная приставка принимает жен. р. («*одну целую*», «*две целые три десятые*»).

Также отдельно обрабатываются обыкновенные дроби, для которых числитель формируется и склоняется как количественное числительное, а знаменатель – как порядковое. Пример обработки обыкновенных дробей показан на рисунке 5.



***Рис. 5 – пример преобразования обыкновенной дроби в числительное***

Отдельно прописано правило: если последним словом числителя не является «*один, два, три, четыре*», а общий падеж выбран именительный или винительный, то падеж знаменателя заменяется на родительный (например, «*пять восьмЫХ*», но «*четыре восьмЫЕ»*).

3.2 Определение верной расшифровки неоднозначных сокращений с помощью контекстных векторов

В качестве базы для обработки сокращений был составлен словарь, совмещающий в себе сгенерированный словарь по материалам вышеописанных корпусов текстов, а также наиболее частотные сокращения из базы данных русских сокращений [RU SOKR]. Итоговый объем ~ 7 тыс. сокращений. Для предсказания леммы сокращения использовался метод соотнесения сокращения с его полной формой в составленном словаре. Особую сложность представляли неоднозначные сокращения, которые в зависимости от контекста могут быть расшифрованы по-разному (ср.: «нет 5г сахара»/шел 5г засухи»). Словарь был составлен таким образом, что одному такому сокращению соответствует набор его возможных значений. Для таких случаев была разработана система снятия неоднозначности сокращения, основанная на тезаурусе RuWordNet и системе контекстуализированного векторного представления контекстов.

*WordNet* — это машиночитаемая лексическая база данных, которая описывает семантические и лексические отношения между словами в английском языке. Она была создана в Принстонском университете в 1985 году, и с тех пор была переведена на несколько языков [Miller, 1995].

WordNet состоит из синсетов (групп слов, имеющих общее значение, наборов когнитивных синонимов) и связей между ними. Каждый синсет содержит набор слов, которые имеют похожее значение. Например, синсет "dog" может содержать слова "puppy", "canine", "hound" и т.д.

Синсеты связаны между собой посредством понятийно-семантических и лексических отношений. Связи между синсетами в WordNet включают гипонимию (отношение "является видом"), гиперонимию (отношение "является родом"), меронимию (отношение "является частью"), холонимию (отношение "является целым"), отношения антонимии, причины, логического следования, предметной области (домена). Кроме того, синсет содержит краткое определение и, в большинстве случаев, одно или несколько коротких предложений, иллюстрирующих использование членов синсета.

*RuWordNet* — это русскоязычная версия лексической базы данных WordNet. Она была создана на основе автоматизированной конвертации тезауруса RuThes [Лукашевич, 2010] в формат WordNet и содержит более 40 000 синсетов и около 200 000 слов. База содержит синсеты трех частей речи: существительных (29297 синсетов), глаголов (7636 синсетов) и прилагательных (12865 синсетов) [Loukachevitch и др., 2016].

RuWordNet может быть использована для создания семантических моделей и анализа текстов на семантическом уровне. Однако, использование данной базы может быть ограничено, так как она не содержит всех слов в русском языке и может содержать ошибки в классификации слов.

Так, синсет RuWordNet может послужить некоторым эталоном для представления значения слова, основываясь на связи с другими лексическими единицами, которые он описывает. Он, конечно, не может дать полные знания об использовании слова в контексте, однако представления о значении и связи с другими словами, которые предоставляются RuWordNet, могут быть использованы для разграничения разных по смыслу слов, имеющих одинаковое сокращенное представление.

Для каждого сокращения в тексте находится его соответствие в словаре сокращений и выбирается полная форма. Если сокращение содержит одно значение (напр., *км* — километр), то данное оно выбирается как лемма. Если сокращение многозначное, то из словаря выбирается список возможных значений. Для каждого из возможных значений составляется псевдо-предложение, которое представляет собой список слов из синсета — синонимов, примеров употребления, определения и других семантически связанных слов, на основе которых и производится выбор между возможными вариантами.

Для совершения выбора псевдо-предложения представляются в виде числового вектора с помощью контекстуализированной модели векторного представления значений.

*Векторное представление* (эмбеддинги, от англ. "embedding") – это процесс представления слов или фраз в виде числовых векторов, которые могут быть использованы для обучения моделей машинного обучения.

Основная идея заключается в том, что каждое слово в тексте представляется в виде вектора в n-мерном пространстве, где n – это размерность эмбеддинга. Эти векторы содержат информацию о семантическом значении слова и его контексте, что позволяет модели машинного обучения лучше понимать естественный язык. Так, более близкие по значению слова будут представлены в векторном пространстве более близкими векторами. Данные системы основываются на дистрибутивной гипотезе, согласно которой близкие по значению слова используются в схожих контекстах [Mikolov и др., 2013a]. Таким образом, в векторном пространстве обученной модели эмбеддингов вектора связаны между собой по семантическому признаку. Это прослеживается на примерах, которые приводятся в исследованиях, посвященных векторным представлениям слов: например, согласно [Mikolov, Yih, Zweig, 2013], если вычесть из векторного представления слова *король* векторное представление слова *мужчина* и прибавить к нему векторное представление слова *женщина*, то получится вектор, наиболее близкий к векторному представлению слова *королева.*

Существуют различные методы создания эмбеддингов, такие как Word2Vec [Mikolov и др., 2013b], GloVe [Pennington, Socher, Manning, 2014], FastText [Joulin и др., 2016] и другие. Они используются в различных задачах обработки естественного языка, таких как классификация текстов, машинный перевод, анализ тональности и др.

*Контекстуализированные модели векторного представления* — это модели эмбеддингов, которые позволяют создавать векторные представления слов и фраз, учитывая контекст, в котором они используются.

Одним из наиболее известных примеров контекстуализированных моделей векторного представления является модель BERT (Bidirectional Encoder Representations from Transformers), разработанная исследователями компании Google [Devlin и др., 2019]. BERT является автокодировщиком: эта модель основана на технологии трансформеров для обработки текста и использует двунаправленную модель внимания, которая анализирует все слова в предложении одновременно в левом и правом контексте.

В настоящей работе используется предобученный вариант модели векторного представления предложений LaBSE (Language-agnostic BERT Sentence Embedding) [Feng и др., 2022]. Данная модель обучена на 109 языках и показывает высокую производительность на разных задачах обработки языка.

Для каждого неоднозначного сокращения контекст его использования представляется в виде вектора, полученного с помощью модели LaBSE. Также в форме векторов представляются и возможные варианты его полной формы, полученные из словаря сокращений. Для выбора наиболее вероятной полной формы сокращения, вычисляется косинусное сходство между вектором контекста и векторами псевдо-предложений, представляющих возможные варианты расшифровки сокращения. Косинусное сходство вычисляется по следующей формуле:

$$similarity = cos(θ) = \frac{A ⋅ B}{||A|| ||B||} = \frac{\sum\_{i = 1}^{n}A\_{i} × B\_{i}}{\sqrt{\sum\_{i = 1}^{n}(A\_{i})^{2} }× \sqrt{\sum\_{i = 1}^{n}(B\_{i})^{2} }}$$

Значение, соответствующее вектору, наиболее близкому к вектору контекста и выбирается как полная форма сокращения.

**4. Модуль предсказания грамматической формы**

Модуль предсказания грамматической формы основан на статистической модели классификации CRF. Использование модели CRF для предсказания грамматической формы русских числительных описывается в работе [Muzychka, Romanenko, Piontkovskaja, 2014].

4.1 Описание модели условных случайных полей

Модель условных случайных полей (Conditional Random Fields, CRF) – дискриминативная статистическая модель классификации, являющаяся разновидностью Марковских случайных полей и учитывающая контекст [Sutton, McCallum, 2006]. Такая модель активно используется в задачах обработки естественного языка, например, для распознавания именованных сущностей [Shahgir и др., 2023]. Задача восстановления полных форм числительных и сокращений сводится к разметке последовательности, где для каждого числительного/сокращения определяются грамматические характеристики, а также вычисляются характеристики для ближайшего контекста. Признаки включают грамматические метки слов, особенности написания и др. Задача формализуется как:

y ∗ = arg max y p(y|x),

где y - последовательность меток грамматических категорий, а x - последовательность наблюдаемых переменных (признаков).

Преимущество использования CRF заключается в том, что такая модель может учитывать контекст, предсказывая метку для одной выборки, учитывая соседние выборки. Для этого прогнозы моделируются в виде графической модели, которая представляет наличие зависимостей между прогнозами.

4.1.1 Формальное определение CRF

Определение CRF взято из работы [Антонова, Соловьев, 2013]. CRF – это тип дискриминационной неориентированной вероятностной графической модели, который можно рассматривать как частный случай Марковского случайного поля. Формально Марковское случайное поле состоит из следующих компонентов:

* неориентированный граф или фактор-граф G = (V, E), где каждая вершина v∈V является случайной переменной Х и каждое ребро {u,v}∈E представляет собой зависимость между случайными величинами u и v.
* набор потенциальных функций (potential function) или факторов {φk}, одна для каждой клики в графе G (полный подграф). Функция φk ставит каждому возможному состоянию элементов клики в соответствие некоторое неотрицательное вещественнозначное число.

Вершины, не являющиеся смежными, должны соответствовать условно независимым случайным величинам. Группа смежных вершин образует клику, набор состояний вершин является аргументом соответствующей потенциальной функции.

Совместное распределение набора случайных величин Х={xk} в Марковском случайном поле вычисляется по формуле (1):

1. $P(x) = \frac{1}{Z}\prod\_{k}^{}φ\_{k}(x\_{\{k\}})$,

где $φ\_{k}(x\_{\{k\}})$— потенциальная функция, описывающая состояние случайных величин в k-ой клике; Z — коэффициент нормализации —вычисляется по формуле (2):

1. $ Z = \sum\_{x\in X}^{}\prod\_{k}^{}φ\_{k}(x\_{\{k\}})$.

Множество входных лексем X={xt } и множество соответствующих им типов Y={yt } в совокупности образуют множество случайных переменных V=X∪Y. Для решения задачи извлечения информации из текста достаточно определить условную вероятность P(Y|X). Потенциальная функция имеет вид:

1. $ φ\_{k}(x\_{\{k\}}) =exp(\sum\_{k}^{}λ\_{k}f\_{k}(y\_{y}, y\_{t-1}, x\_{t}))$,

где Σ{λk} вещественнозначный параметрический вектор, и Σ{fk(yt, yt-1, xt)} — набор признаковых функций. Тогдалинейным условным случайным полем называется распределение вероятности вида:

1. $ p(x|y)=\frac{1}{z(x)}\prod\_{k}^{}exp(\sum\_{k}^{}λ\_{k}f\_{k}(y\_{y}, y\_{t-1}, x\_{t}))$;

тогда коэффициент нормализации Z(x) вычисляется по формуле (5):

1. $ Z(x) = \sum\_{y}^{}\prod\_{k}^{}exp(\sum\_{k}^{}λ\_{k}f\_{k}(y\_{y}, y\_{t-1}, x\_{t}))$.

Линейно-цепочечное УСП (linear-chain CRF) — это условное случайное поле, у которого множество скрытых переменных вытянуто в цепочку.

4.2 Применение модели для предсказания формы числительных

В процессе эксперимента была реализована модель линейно-цепочечной CRF, предсказывающая грамматические категории числительных, представленных в тексте в виде чисел.

Для обучения модели из корпуса были выбраны числительные, содержащие морфологические теги в разметке. В связи с тем, что в данных корпусах морфологические теги были приписаны только числительным, данные числительные заменялись на их числовую запись для формирования обучающих данных. Итоговое количество предложений, содержащих числительные составило 49171. Теги разных корпусов были приведены к общему стандарту UPOS (от англ. Universal Part-of-Speech).

В связи с тем, что в представленных в корпусе примерах содержались часто употребляемые конструкции, содержащие числа, которые крайне редко представлены в орфографическом виде (например, конструкции с лексемой *год*, такие как *в 1675 году*, *за 2019-ый год* и другие) было принято решение провести ручную доразметку данных на основе лексем, часто употребляемых с числительными. Из корпуса были выбраны предложения, содержащие лексемы *год,* названия месяцев и валют, перед которыми употреблялось число. Числительным в этих предложениях был вручную приписан грамматический тег. Объем доразмеченных данных составил 5654 контекста.

Каждому числительному была приписана метка класса, содержащая следующие грамматические категории:

* количественное/порядковое,
* падеж,
* род,
* число,
* одушевленность.

Если числительному несвойственна какая-то из грамматических категорий, то она заменялась 0 (например, род для количественных числительных больше двух). Остальным частям речи приписывался тег “О”.

Кроме того, был определен список словоформ, которые требуют использования перед ними числительного определенного типа (порядкового или количественного). К таким словам относятся, например, названия валют и месяцев.

Каждый токен контекста был представлен в виде вектора признаков:

* словоформа,
* три последние символа словоформы,
* два последние символа словоформы,
* является ли токен цифрой,
* частеречный тег,
* является ли токен первым словом в предложении,
* является ли токен последним словом в предложении,
* является ли токен словом, требующим использования перед собой определенного типа числительного (порядкового или количественного);
* а также характеристики двух предыдущих и двух последующих слов. К таким характеристикам относятся:
	+ словоформа,
	+ длина слова,
	+ три последних символа словоформы,
	+ два последних символа словоформы,
	+ является ли токен цифрой,
	+ начинается ли токен с заглавной буквы,
	+ частеречный тег,
	+ является ли токен словом, требующим использования перед собой определенного типа числительного (порядкового или количественного).

Данные признаки были выбраны на основе работы [Романенко, 2014]. Некоторые признаки были добавлены дополнительно, например информация о двух и трех последних символов слова. Данная информация в большинстве случаев соответствует флексии, по которой можно судить о грамматической форме ближайших слов. Некоторые признаки были добавлены для получения дополнительной статистической информации моделью во время обучения — среди таковых, например - информация о том, является ли токен первым или последним словом в предложении. На основе этой информации модель может делать выбор в спорных ситуациях, основываясь на том, числительные в какой форме чаще стоят в начале или конце предложений.

Данные были разделены на тестовую и обучающую выборки. Размер обучающей выборки составил 80% всех примеров, остальная часть использовалась для теста.

Для реализации алгоритма использовалась python библиотека sklearn\_crfsuite.

4.3 Модель CRF для определения грамматической формы сокращений

Для обучения модели распознавания грамматической формы сокращений из корпуса были отобраны собственно сокращения (исключались аббревиатуры, инициалы, сокращенные формы числительных, такие как “80-х”, которые в корпусе помечены тем же тегом *Abbr*, что и сокращения). Всего было выделено 7030 вхождений сокращений, среди которых 120 уникальных токенов. Все сокращения, представленные в корпусе, имеют полную морфологическую разметку.

В связи с непоследовательной маркировкой элементов корпуса тегом *Abbr* и ограниченной представленностью сокращений в корпусе текста, было принято решение провести аугментацию данных. Аугментация основывалась на словаре сокращений. Для каждого вхождения слова из корпуса, для которого имеется сокращение в данном словаре, полная форма этого слова была заменена на его сокращение. Например: *“какая цена за квадратный метр?” – “Какая цена за кв.м.?”*. Таким образом, количество обучающих данных было увеличено до 20000 вхождений сокращений.

Каждому сокращению приписывался тег, соответствующий его грамматической форме. Остальным частям речи приписывался тег “О”.

Каждый токен контекста был представлен в виде вектора признаков:

* словоформа,
* длина слова,
* три последние символа словоформы,
* два последние символа словоформы,
* является ли токен знаком препинания,
* частеречный тег,
* является ли токен первым словом в предложении,
* является ли токен последним словом в предложении,
* а также характеристики двух предыдущих и двух последующих слов. К таким характеристикам относятся:
	+ словоформа,
	+ длина слова,
	+ три последних символа словоформы,
	+ два последних символа словоформы,
	+ является ли токен цифрой,
	+ начинается ли токен с заглавной буквы, частеречный тег.

Данные признаки также выделялись на основе работы [Романенко, 2014], соответственно, без использования признаков, необходимых только для числительных: является ли токен цифрой и является ли токен словом, требующим использования перед собой определенного типа числительного.

На основе этих данных делалось предсказание. Данные были разделены на обучающий и тестовый наборы, тестовый набор составил 20% от всей выборки.

**5. Результаты**

В качестве основной меры для оценки качества модели использовалась F1 мера, которая вычисляется по формуле (6):

1. $F1 =\frac{2TP}{2TP + FP + FN} $,

где TP — количество истинно положительных срабатываний классификатора, FP — количество ложно положительных и FN — количество ложно отрицательных срабатываний [Taha, Hanbury, 2015].

5.1 Оценка работы классификатора для предсказания формы числительного

Модель для предсказания грамматической формы числительных показала значение F1-меры равное 0,86 на тестовых данных. Результаты представлены в таблице 1.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Точность (precision) | Полнота (recall) | F1-мера |
| micro avg | 0.86 | 0.86 | 0.86 |
| macro avg | 0.56 | 0.52 | 0.52 |
| weighted avg | 0.87 | 0.86 | 0.86 |

*Таблица 1 – результаты предсказания грамматической формы числительных*

На основе результатов тестирования модели был проведен анализ ошибок, возникших в процессе предсказания. По результатам анализа были выявлены группы контекстов, ошибки в которых могут отрицательно влиять на точность модели, при этом не влияя на ее работоспособность. Это такие ошибки как, например:

1. невозможность точного определения грамматической формы целевого числительного в связи с недостатком контекста. К таким примерам относятся некоторые предложения корпуса, попавшие в тестовую выборку, например: *“с надписью уценка и что же 3 четырехмесячной давности караван историй свежий”*, *“важен рассказ геродота herod 64 83 о празднике деметры”*. В таких случаях даже человек не может корректно определить форму числительного в связи с тем, что контекста не хватает для понимания смысла.
2. неинформационные контексты. Корпус текстов, использованный в данном эксперименте для обучения и тестирования сети, довольно большой, поэтому иногда в нем встречаются контексты, которые мало похожи на настоящие тексты, например “*so 16 so 19 so 20 so 95 so 2 so*”, “*52 6 это в среднем 400 этого 49 10 при этом*”.
3. некорректные контексты. В связи с тем, что корпус включает в себя тексты разных типов и жанров, в некоторых текстах содержатся ошибки. Например, “*теперь тоже было 24 игрушек*” (для того, чтобы предложение было корректным, нужно использовать вариант “*теперь тоже было 24 игрушки*”).
4. некорректно размеченные числительные. В обучающих данных встречаются неправильные грамматические метки для некоторых числительных. Например, “*операция была проведена в америке в 1842 году*”. В данном предложении токену *1842* модель CRF предсказывает метку, которая соответствует корректной в данном контексте форме порядкового числительного в предложном падеже, единственного числа, мужского рода, в то время как в корпусе данный токен размечен как количественное числительное мужского рода единственного числа в именительном падеже.

Кроме таких ошибок, связанных с качеством корпусных данных, есть случаи, когда грамматически теги приписываются некорректно (то есть, фактически, модель совершает ошибку в предсказании), однако предсказанная форма совпадает с целевой формой в плане выражения. В пример можно привести несколько разных предложений, например, *“так как члены 2 групп осознают”*. В разметке токену 2 приписан тег *“количественное числительное женского рода в родительном падеже”* (т.е. соответствующий форме *“двух”*), в то время как предсказание модели — тег *“количественное числительное мужского рода, одушевленное, в винительном падеже”* (который также соответствует форме *“двух”*). Таким образом, несмотря на некорректное предсказание модели, итоговая нормализация текста будет верной.

Однако, модель совершает и такие ошибки, которые нельзя объяснить некорректностью данных. Нами был проведен анализ таких контекстов и были выявлены наиболее частые конструкции, в которых предсказание оказывается некорректным:

1. конструкции, в которых опущено слово, с которым согласуется числительное. Например, “*Он открыл счёт с пенальти на 35-й. Снова Лопес остановил мяч*”. Слово “*минуте*” после числительного опущено в контексте, что свойственно для спортивных комментариев. Без возможности опираться на управляющее существительное, грамматическая метка предсказывается некорректно — в данном случае была предсказана форма “*порядковое числительное, именительный падеж, мужской род, единственное число*” (метка соответствует форме “*тридцать пятый*”).
2. конструкции, где слово, с которым согласуется числительное, находится на большом расстоянии от числительного. Например, в контексте “*1-ое жило в ирландском городе, а 2 в окружённом стеной английском*” слово, на основе которого можно сделать вывод о роде порядкового числительного, находится на расстоянии пяти токенов, поэтому в таких случаях модель может ошибаться. Об этом типе ошибок также сообщалось в работе [Хомицевич, Рыбин, Аничкин, 2013]. Для решения требуется синтаксический анализ более отдаленных связей.
3. конструкции, где ключевыми для определения формы являются синтаксические связи. Например, “*будут пущены в дорогу 6 библиомобилей*”. Поскольку в данном примере от типа связи зависит падеж числительного (именительный в случае управления «*шесть библиомобилей*» и родительный в случае согласования «*шести библиомобилей*»), данная проблема может быть решена на синтаксическом уровне путем добавления к модели синтаксических признаков.

5.2 Оценка работы классификатора для предсказания формы сокращения

Модель для предсказания грамматической формы сокращений показала значение F1-меры равное 0,88 на тестовых данных. Результаты представлены в таблице 2.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Точность (precision) | Полнота (recall) | F1-мера |
| micro avg | 0.88 | 0.87 | 0.88 |
| macro avg | 0.72 | 0.67 | 0.69 |
| weighted avg | 0.88 | 0.87 | 0.87 |

*Таблица 2 – результаты предсказания грамматической формы сокращений*

Результаты тестирования были проанализированы с целью выявления групп контекстов, в которых ошибки встречаются наиболее часто. В случае с сокращениями также имеют место ошибки, возникающие в связи с неудовлетворительным качеством данных. Например, в предложении “*он находился на отметке 797 см.*” сокращение “*см.*” размечено в корпусе как глагол повелительного наклонения. Судя по всему, разметчик подразумевал полную форму “*смотрите*”, однако в данном контексте корректная расшифровка “*сантиметр*”. Так, модель предсказывает корректные грамматические теги, которые засчитываются как ошибочные в связи с неправильной разметкой и недостаточным контекстом. Большая часть ошибок возникает в связи с недостатком контекста, так как выражения, составляющие тестовую выборку, ограничены по длине, в связи с чем часть выражения может обрываться. Кроме того, разметка данных, которые были в корпусе изначально, а не добавлялись средствами аугментации, содержит очень большое количество ошибок. Так, например, “*они включают 35 млн* (размечено как форма именительного падежа) *замен отдельных нуклеотидов*”, “*оговоренные в документе 7,2 млн* (размечено как форма творительного падежа) *долларов золотом не были оправлены*”, “*мемориал кирову на кировской пл.”* (размечено как форма именительного падежа) и пр.

Наиболее часто ошибки, не относящиеся к качеству данных, случаются в следующих случаях:

1. cокращение находится на большом расстоянии от слова, с которым оно согласуется;
2. перед сокращением присутствует ненормализованное числительное. Таких примеров в корпусе достаточно много, например, предложение “*Призовой фонд составил не менее 1 млн. долларов США*”. В таких случаях модель часто предсказывает начальную форму сокращения (в данном случае форму единственного числа именительного падежа). Такие ошибки нивелируются созданием параллельных корпусов, где всем предложениям соответствуют нормализованные варианты.

Помимо этих групп встречаются также ошибки, наличие которых можно объяснить только недостаточной представленностью данной формы, выражения или конструкции в обучающих данных, в связи с чем корректное предсказание становится невозможным.

ЗАКЛЮЧЕНИЕ

В ходе настоящего исследования были проанализированы подходы к нормализации текста на разных языках, а затем на основе анализа выбран оптимальный подход для русского языка. Также было рассмотрено место нормализации числительных и сокращений в системах синтеза речи и исследованы числительные и сокращения как функционирующие языковые единицы.

В рамках практической главы была предложена система нормализации текста, основанная на комбинации формальных правил и статистической модели условных случайных полей. Эта система была протестирована и оценена отдельно для числительных и сокращений.

По результатам оценки предложенной системы были получены следующие выводы:

* + - 1. задача разрешения лексической неоднозначности расшифровок сокращений - пример высокоуровневой проблемы семантической классификации, нуждающейся в наборе признаков, который невозможно легко вывести из поверхностных форм. Для ее решения полезна лингвистическая предобработка на синтаксическом уровне, а также получение информации из поддерживаемых вручную семантических лексиконов.
			2. нейросетевые подходы достигли высокого уровня точности, однако они имеют тенденцию к совершению критических ошибок, в частности, неверно определяют основу числительного ("*3-х*" как "*пять тысяч*" вместо "*трёх*"). Несмотря на крайне малый процент данных ошибок, они носят спонтанный характер, поэтому невозможно предугадать, когда система «исказит» фактические данные, которые должны быть точны.
			3. глубокое обучение требует значительного количества вычислительных ресурсов, обучающих данных и временных затрат. В этой связи комбинация формальных правил и статистического классификатора CRF предлагает более надежное решение, которое не настолько трудо- и ресурсоемко. Предложенный метод показывает приемлемую точность предсказания грамматической формы (86%), более устойчив к критическим ошибкам при определении основы и более прост в условиях эксплуатации в реальной системе синтеза речи по тексту.

Таким образом, результаты исследования подтверждают, что предложенная система нормализации текста на основе комбинации формальных правил и статистической модели CRF является эффективным и надежным решением для русского языка.

Полученные результаты и выводы могут служить основой для дальнейших исследований в области нормализации текста и развития более продвинутых систем нормализации нестандартных записей в системах синтеза речи для флективных языков.

Список литературы

1. Аничкин И. М., Чистиков П. Г. Формализация правил автоматического снятия омонимии в системе синтеза речи по тексту // Труды XXXVIII международной филологической конференции.–2008.–C. , 2008. С. 29–45.

2. Антонова А. Ю., Соловьев А. Н. Использование метода условных случайных полей для обработки текстов на русском языке // Диалог. 2013. С. 27–44.

3. Брызгунова Е. А. и др. Русская грамматика: научные труды/Российская академия наук. Институт русского языка им. ВВ Виноградова/. Репринтное издание // 2005.

4. ГОСТ Р. 7.0. 12—2011 Библиографическая запись. Сокращение слов и словосочетаний на русском языке. Общие требования и правила // М.: Гостстандарт. 2011. Т. 201. № 1. С. 42.

5. Зализняк А. «Русское именное словоизменение» с приложением избранных работ по современному русскому языку и общему языкознанию. : Litres, 2022.

6. Зализняк А. А. Грамматический словарь русского языка. : Русские словари, 2003.

7. Иванова В. Современный русский язык: графика и орфография:[учебное пособие]. : Просвещение, 1976.

8. Лобанов Б. М. Методы семантического анализа для построения голосовых интерфейсов: синтез речи // 2011.

9. Лобанов Б. М., Цирульник Л. И. Компьютерный синтез и клонирование речи // Минск: Белорусская наука. 2008. Т. 316.

10. Лукашевич Н. В. Тезаурусы в задачах информационного поиска // 2010.

11. Романенко А. А. Применение условных случайных полей в задачах обработки текстов на естественном языке // Выпускная квалификационная работа магистра, М. 2014.

12. Рыбин С. В. Синтез речи: учебное пособие // СПб.: Университет ИТМО. 2014.

13. Санников В. З. О русских графических сокращениях // О современной русской орфографии/Отв. ред. ВВ Виноградов. М. 1964.

14. Хомицевич О. Г., Рыбин С. В., Аничкин И. М. Использование лингвистического анализа для нормализации текста и снятия омонимии в системе синтеза русской речи // Известия высших учебных заведений. Приборостроение. 2013. Т. 56. № 2. С. 42–46.

15. Allen J. и др. From text to speech: The MITalk system. : Cambridge University Press, 1987.

16. Arık S. Ö. и др. Deep voice: Real-time neural text-to-speech // International conference on machine learning. : PMLR, 2017. С. 195–204.

17. Beliga S., Martincic-Ipsic S. Non-standard words as features for text categorization // 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). : IEEE, 2014. С. 1165–1169.

18. Black A. W., Taylor P., Caley R. The Festival Speech Synthesis System: System Documentation Festival version 1.4 // University of Edinburgh. 1999. Т. 1.

19. Breuillard J. Terence Wade, A Comprehensive Russian Grammar, 1992 // La Revue russe. 1992. Т. 3. № 1. С. 109–115.

20. Chan W. и др. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition // 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). : IEEE, 2016. С. 4960–4964.

21. Chu M., Qian Y. Locating boundaries for prosodic constituents in unrestricted Mandarin texts // International Journal of Computational Linguistics & Chinese Language Processing, Volume 6, Number 1, February 2001: Special Issue on Natural Language Processing Researches in MSRA. , 2001. С. 61–82.

22. Corstius H. B. Grammars for number names. : Springer Science & Business Media, 2012.

23. Devlin J. и др. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // 2019.

24. Dutoit T. An introduction to text-to-speech synthesis. : Springer Science & Business Media, 1997.

25. Epstein S. James R. Hurford, The linguistic theory of numerals.(Cambridge Studies in Linguistics, 16.) Cambridge: Cambridge University Press, 1975, Pp. xii+ 293. J. Linguistics 14 (1978) 123–127 Printed in Great Britain // Journal of Linguistics. 1978. Т. 14. № 1. С. 123–124.

26. Feng F. и др. Language-agnostic BERT Sentence Embedding // 2022.

27. Gorman K., Sproat R. Minimally supervised number normalization // Transactions of the Association for Computational Linguistics. 2016. Т. 4. С. 507–519.

28. Joulin A. и др. Bag of Tricks for Efficient Text Classification // 2016.

29. Jurafsky D., Martin J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.

30. Lemmetty S. Review of speech synthesis technology // 1999.

31. Loukachevitch N. V. и др. Creating Russian wordnet by conversion // Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue. , 2016. С. 405–415.

32. Lukanin A. Normalization of non-standard words with finite state transducers for russian speech synthesis // Analysis of Images, Social Networks and Texts: 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers 4. : Springer, 2015. С. 39–48.

33. Mansfield C. и др. Neural text normalization with subword units // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers). , 2019. С. 190–196.

34. Mikolov T. и др. Efficient Estimation of Word Representations in Vector Space // 2013a.

35. Mikolov T. и др. Distributed Representations of Words and Phrases and their Compositionality // Advances in Neural Information Processing Systems. : Curran Associates, Inc., 2013b.

36. Mikolov T., Yih W., Zweig G. Linguistic Regularities in Continuous Space Word Representations // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia: Association for Computational Linguistics, 2013. С. 746–751.

37. Miller G. A. WordNet: a lexical database for English // Communications of the ACM. 1995. Т. 38. № 11. С. 39–41.

38. Muzychka S. A., Romanenko A. A., Piontkovskaja I. I. Conditional Random Field for morphological disambiguation in Russian // Компьютерная лингвистика и интеллектуальные технологии. , 2014. С. 455–465.

39. Ng A. H., Gorman K., Sproat R. Minimally supervised written-to-spoken text normalization // 2017 IEEE automatic speech recognition and understanding workshop (ASRU). : IEEE, 2017. С. 665–670.

40. Pennington J., Socher R., Manning C. Glove: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014. С. 1532–1543.

41. Shahgir H. A. Z. и др. Towards Robust Bangla Complex Named Entity Recognition // arXiv preprint arXiv:2303.09306. 2023.

42. Shen Q. и др. The role of context in neural morphological disambiguation // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. , 2016. С. 181–191.

43. Sproat R. и др. Normalization of non-standard words // Computer speech & language. 2001. Т. 15. № 3. С. 287–333.

44. Sproat R. Lightly supervised learning of text normalization: Russian number names // 2010 IEEE Spoken Language Technology Workshop. : IEEE, 2010. С. 436–441.

45. Sproat R., Jaitly N. RNN approaches to text normalization: A challenge // arXiv preprint arXiv:1611.00068. 2016.

46. Sproat R. W. Multilingual text-to-speech synthesis: the Bell Labs approach. : Springer, 1998.

47. Stanton D., Wang Y., Skerry-Ryan R. J. Predicting expressive speaking style from text in end-to-end speech synthesis // 2018 IEEE Spoken Language Technology Workshop (SLT). : IEEE, 2018. С. 595–602.

48. Sutton C., McCallum A. An introduction to conditional random fields for relational learning. Introduction to statistical relational learning // Graphical Models,(x). 2006.

49. Taha A. A., Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool // BMC Med Imaging. 2015. Т. 15. С. 29.

50. Tan X. и др. A survey on neural speech synthesis // arXiv preprint arXiv:2106.15561. 2021.

51. Taylor P. Text-to-speech synthesis. : Cambridge university press, 2009.

52. Yao K., Zweig G. Sequence-to-sequence neural net models for grapheme-to-phoneme conversion // arXiv preprint arXiv:1506.00196. 2015.

53. Yolchuyeva S., Németh G., Gyires-Tóth B. Text normalization with convolutional neural networks // International Journal of Speech Technology. 2018. Т. 21. С. 589–600.

54. Żelasko P. Expanding Abbreviations in a Strongly Inflected Language: Are Morphosyntactic Tags Sufficient? // arXiv preprint arXiv:1708.05992. 2017.

55. Zelinka J., Romportl J., Müller L. Statistical-based abbreviation expansion // Text, Speech and Dialogue: 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings 14. : Springer, 2011. С. 307–314.

56. Zhang H. и др. Neural models of text normalization for speech applications // Computational Linguistics. 2019. Т. 45. № 2. С. 293–337.

57. Zhang J. и др. A hybrid text normalization system using multi-head self-attention for mandarin // ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). : IEEE, 2020. С. 6694–6698.

58. Открытый корпус русских текстов [Электронный ресурс]. URL: http://www.opencorpora.org/ (дата обращения: 30.04.2023).

59. Национальный корпус русского языка. 2003-2023. [Электронный ресурс]. URL: https://ruscorpora.ru/ (дата обращения: 20.05.2023).

60. RU SOKR. самый большой в мире словарь сокращений русского языка [Электронный ресурс]. URL: http://www. sokr. ru/card/233514s38737f58 (дата обращения: 01.04.2023).