УДК 81’33:004.822

Автоматизация построения терминологического ядра онтологии по компьютерной лингвистике на основе корпуса текстов

Овчинникова К. А. (Новосибирский национальный исследовательский государственный университет),

Иванов А. И. (Новосибирский национальный исследовательский государственный университет),

Сидорова Е. А. (Институт систем информатики СО РАН)

В работе предлагается подход к автоматическому построению терминологического ядра онтологии по компьютерной лингвистике. Рассматриваются вопросы создания онтологии верхнего уровня, определяющей возможные классы терминов для их дальнейшего поиска и систематизации. Предложен алгоритм генерации и начального пополнения предметного словаря, включающий два основных этапа. На первом шаге строится система лексико-семантических классов, основанных на классах онтологии. На втором шаге осуществляется наполнение словаря терминами и их соотнесение с классами словаря на основе имеющихся ресурсов: универсальной онтологии научного знания, тезауруса и портала по компьютерной лингвистике. Для проведения экспериментального исследования был собран корпус аналитических статей по компьютерной лингвистике с сайта Хабр и созданы наборы данных с разметкой терминов, включающие по 1065 предложений на русском языке. Проведены эксперименты для решения двух задач: обнаружение терминов и их классификация относительно классов онтологии. Для первой задачи были рассмотрены три нейросетевые модели: xlm-roberta-base, roberta-base-russian-v0 и ruRoberta-large. Лучшие результаты получены на последней модели: 0.91 F-меры. Проведен анализ ошибок классификатора, который показал, что высокую частотность ошибки неполного выделения термина. Для второй задачи была выбрана модель ruRoberta-large, показавшая лучшие результаты для первой задачи. Среднее значение F-меры для 12 используемых классов онтологии составило 0.89. Предложена общая архитектура системы создания и пополнения онтологий, интегрирующая лингвистические подходы и методы машинного обучения.

Ключевые слова: *терминологическое ядро онтологии; компьютерная лингвистика; онтология компьютерной лингвистики; извлечение терминов; классификация терминов.*

1. Введение

Систематизация знаний в активно развивающейся области, такой как компьютерная лингвистика (КЛ), является важной, но ресурсоемкой задачей. Большое количество информационных ресурсов, приложений, моделей требуют оперативного решения задач поиска и анализа информации о последних достижениях науки. Так, например, интерес к применению нейронных сетей для решения задачи автоматической обработки текстов вызвал появление огромного количества новых методов, наборов данных и языковых моделей, знания о которых необходимы специалистам. Для решения этой проблемы разрабатываются различные интернет-ресурсы: каталоги, вики-ресурсы, тематические форумы, порталы знаний [20]. Для описания и систематизации информации используют иерархические модели знаний, в первую очередь графы знаний и онтологии. На основе таких моделей далее создаются решения под управлением знаниями (knowledge-driven applications), которые решают проблему информационной совместимости и формализации и обеспечивают постоянную генерацию новых знаний, непрерывно анализируя множество разрозненных источников информации.

Согласно общепринятому определению в компьютерных науках, *онтология* — это способ формализации знаний, абстрактных или специфических, в какой-либо предметной области, реализованный на основе формального описания объектов, фактов и отношений между ними, и ориентированный на многократное использование для различных задач [27]. Для графов знаний онтология — это семантическая основа представления данных, базирующаяся на логике и включающая терминологический словарь и набор утверждений о моделируемых объектах.

Для разработки новых онтологий могут быть использованы так называемые *онтологии верхнего уровня (*или *базовые онтологии)*, которые включают основные или базовые понятия и отношения, используемые для описания и формализации знаний в целом классе предметных областях, объединенных общим назначением, например, научные предметные области. Так, в качестве основы для описания научной области может быть использована универсальная онтология научной области знаний, представленная в работе [15].

Для конкретизации онтологии на точную предметную область в первую очередь необходимо построить *терминологическое ядро онтологии* — терминологическую систему, описывающую концептуальный состав (множество понятий) онтологии предметной области.

Терминологическая система — это иерархически-структурированная совокупность терминов, принадлежащих к определенной предметной области. Под терминами понимаются слова и словосочетания, являющиеся наименованиями понятий моделируемой области знаний.

Разработка онтологий является сложным и трудоемким процессом, поэтому существует необходимость создания подходов, которые поддерживают и автоматизируют этот процесс. Одним из естественных способов выделить начальный набор понятий (терминов) — обеспечить извлечение названий сущностей предметной области из текстов. Данная задача сводится к двум подзадачам: извлечение ключевых терминов и их классификация относительно классов базовой онтологии. Эту задачу можно отнести к задаче *распознавания именованных сущностей* (named entity recognition, NER), когда в тексте необходимо выделять спаны (непрерывные фрагмента текста) сущностей из заранее заданного набора категорий. Большое количество исследований, посвященных NER, показывает актуальность этой задачи. При этом методы глубокого обучения получают наибольшее внимание, так как часто дают самые лучшие результаты. Однако соревнования, проводимые в данной области демонстрируют, что для разных задач одни и те же методы NER показывают различное качество работы [11].

Целью данного исследования является разработка методов автоматизации построения терминологического ядра онтологии по компьютерной лингвистике на основе базовой онтологии и корпуса текстов. Теоретическая часть работы включала конкретизацию базовой онтологии на область компьютерной лингвистики — требовалось выявить отсутствующие понятия и уточнить основные отношения. Практическая часть работы заключалась в апробации различных методов извлечения терминов и генерации предметных словарей.

Для достижения поставленной цели был собран русскоязычный корпус аналитических статей с веб-сайта Хабр (https://habr.com). С сайта были отобраны статьи, связанные с направлением Natural Language Processing за период с января 2010 года по сентябрь 2023 года (1681 статья).

2. Обзор существующих методов

На данный момент существует различные методы и подходы к построению онтологий. В основном применяются различные методики ручного проектирования онтологии и методы автоматизации отдельных этапов создания онтологии, например, извлечение терминов из текста, классификация терминов, извлечение объектов и отношений и др.

Первые методологии онтологического проектирования предложены еще в конце 90-х годов: Methontology [4], On-To-Knowledge [13], Enterprise Model [14]. Данные подходы были нацелены на создание онтологии предметной области для решения конкретных задач, например, извлечения знаний, систематизации, формализации, поиска информации. На протяжении последних пятнадцати лет развивается подход к разработке онтологий, базирующийся на применении паттернов онтологического проектирования (Ontology Design Patterns или ODP) [5], являющихся стандартизованными описаниями ранее созданных фрагментов онтологий. Суть предлагаемых методик сводится к описанию различных способов переиспользования паттернов при разработке новых онтологий.

Для разработки онтологии по компьютерной лингвистике полезно использовать схожие по тематике ресурсы. На настоящий момент ресурсов по компьютерной лингвистике немного. Так, например, существует русско-английский тезаурус по компьютерной лингвистике [26] и портал по компьютерной лингвистике [25], разработанные в 2010-х годах и не пополняющиеся на данный момент. Также существует онтология Машинного обучения [7], для которой авторы только собираются определить необходимые инструменты для пополнения.

Для автоматизации построения и пополнения онтологий на основе текстов на естественном языке применяются как лингвистические методы, так и методы на основе машинного обучения. К лингвистическим методам можно отнести методы на основе лексико-синтаксических паттернов онтологического проектирования [10, 3]. Они задают отображения языковых структур в онтологические структуры, с помощью шаблонов [19].

Для извлечения терминов и их классификации чаще всего используются методы машинного обучения. Одними из самых ранних подходов для распознавания именованных сущностей были подходы, основанные на классических методах машинного обучения, таких как метод опорных векторов (SVM) или метод случайного леса (Random Forests). Эти методы опираются на признаки слов, такие как принадлежность к синтаксической группе, семантический тип, чтобы в конечном счете определять расстояния между словами и разделить их на классы [16]. Лучшие результаты в данной категории показывают ансамблевые методы.

На сегодняшний день в большинстве случаев используют методы глубокого обучения или гибридные подходы на их основе, о чем свидетельствуют различные обзоры [23, 9]. Для задачи NER в основном используются таких нейросетевые архитектуры как рекуррентные нейронные сети (Recurrent Neural Networks, RNN), рекуррентные нейронные сети с долгой краткосрочной памятью (Long short-term memory, LSTM) или трансформеры типа BERT. Так,

в работе [6] используется разновидность рекуррентных нейронных сетей RNN-T, которая позволяет не только распознавать сущности, но и учитывать их вложенность друг в друга. NER в условиях ограниченных наборов обучающих данных можно рассматривать как отдельную разновидность задачи выделения именованных сущностей. Для решения данной проблемы в работе [17] авторы создают большой набора данных с размеченными именованными сущностями относительно хорошего качества, после чего обучают базовую модель с трансформерной архитектурой NER-BERT на данном наборе данных. После этого авторы обучают базовую модель для более специфичной задачи и более узкого домена сущностей. Этот метод дает лучшие результаты в сравнении с классическими подходами, в которых базовая модель обучается на более общей задаче, такой как маскирование. Созданный набор данных включает только тексты на английском языке, что не дает возможности использовать модель для задач анализа русскоязычного текста. Русскоязычные датасеты достаточного большого объема и качества существуют только для ограниченного набора сущностей, что также ограничивает их применения для более специфичной задачи извлечения терминов по компьютерной лингвистике.

3. Онтология верхнего уровня

В качестве онтологии верхнего уровня для создания онтологии по компьютерной лингвистике была взята онтология, предложенная в [15]. Данная онтология делится на онтологию научной деятельности и онтологию научного знания [21]. *Онтология научной деятельности* включает базовые понятия, относящиеся к организации научно-исследовательской деятельности, такие как *Персона, Организация, Деятельность, Публикация*. *Онтология научного знания* содержит метапонятия и отношения, задающие структуры для описания предметной области (научной дисциплины) портала знаний, например, *Раздел науки, Предмет исследования, Объект исследования, Метод исследования*, позволяющие выделить в данной науке значимые разделы и подразделы, задать типизацию предметов, объектов и методов исследования, описать результаты научной деятельности. Всего универсальная онтология содержит 11 классов и 92 отношения.

Для конкретизации онтологии на область компьютерной лингвистики использовались найденные систематизированные ресурсы: русско-английский тезаурус по компьютерной лингвистике [26] и портал по компьютерной лингвистике [25]. Информация о базовых сущностях предметной области была собрана на основе корпуса с помощью анализа употреблений сущностей в тексте, построения конкордансов и списка частотных терминов.

Полученную онтологию так же, как и базовую, можно условно разделить на две части: то, что относится к научному знанию (Рис. 1, Рис. 2) и то, что можно отнести к научной деятельности (Рис. 3). Ниже описаны особенности, полученные при конкретизации базовой онтологии.

A) Изменена структура онтологии верхнего уровня: убраны классы *Событие* и *Географическое место*, так как предполагается, что онтология будет в большей степени представлять научное знание, а не научную деятельность, несмотря на то, что одно трудно отделяется от другого. Класс *Предмет исследования* был выделен как подкласс класса *Объект исследования*, поскольку предмет исследования в одном исследовании может быть объектом исследования в другом.

B) *Метод исследования* в нашем исследовании является ключевым классом. Современные методы КЛ основаны на методах машинного обучения (МО), поэтому потребовалось ввести такое понятие как *Метод машинного обучения*, который был добавлен как подкласс *Метода исследования* (Рис. 1)

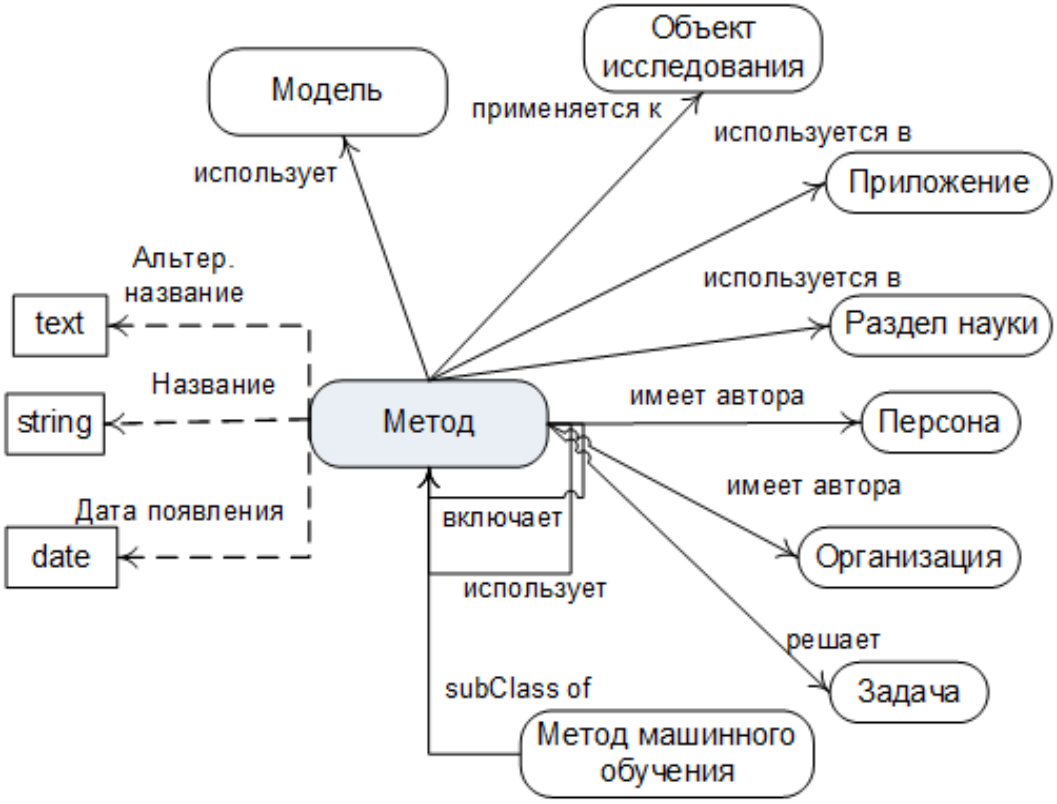

Рис. 1. Фрагмент онтологии, описывающий класс Метода исследования

C) Появление подкласса *Метод Машинного обучения* подразумевают наличие таких классов, как *Модель машинного обучения*, *Набор данных*, *Метрика* (Рис. 2). Выделение класса *Метрика* подтверждается вхождением одноименного термина в корпус текстов 510 раз.

D) Для класса *Модель машинного обучения*, была добавлена аксиома: <*Абстрактная модель* не может иметь связь с *Набором данных*>, т.к. это свойство есть только у конкретной модели.

E) Были добавлены такие классы как *Приложение* и *Окружение* (Рис. 3). Для объектов класса *Приложение* характерны текстовые фрагменты вида:

«в статье использовалась c++ библиотека openfst»

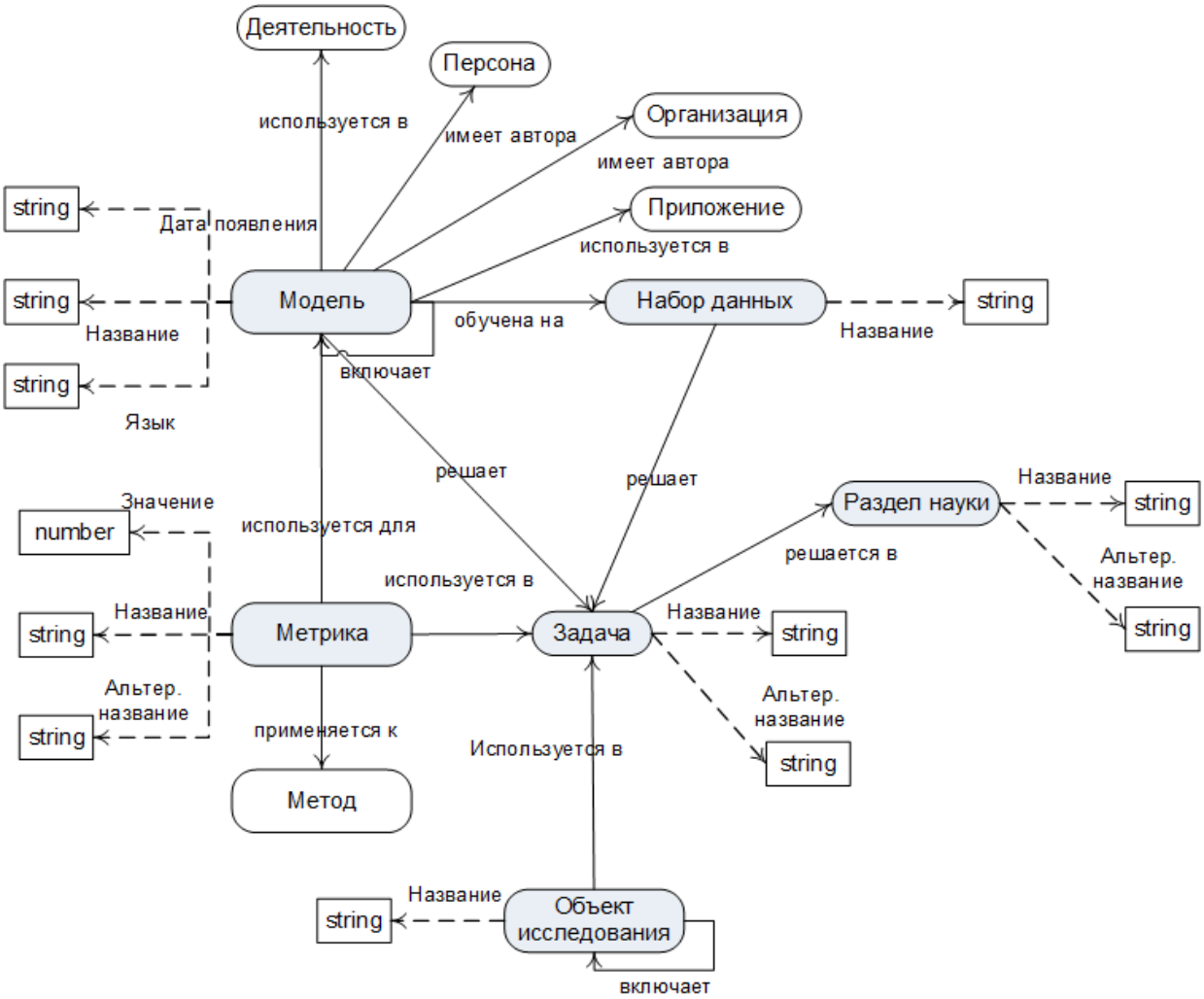

Рис. 2. Фрагмент онтологии, описывающий классы, связанные с научным знанием

F) Классу *Деятельность* было добавлено два важных подкласса, которых не было ранее: *Исследование* и *Эксперимент*. Для объектов класса *Исследование* характерны предложения следующего типа:

«Systran проводит исследование, в котором качество перевода оценивается путем <…>».

Необходимость выделения класса *Эксперимент* подтверждается предложениями типа:

«В рамках Джорджтаунского эксперимента демонстрировалась система, которая автоматически перевела 60 предложений с русского языка на французский».

Добавление класса *Эксперимент* как подкласса класса *Исследование* происходит на основе разделения исследований на два типа: теоретических и практических. Последние мы считаем экспериментами.

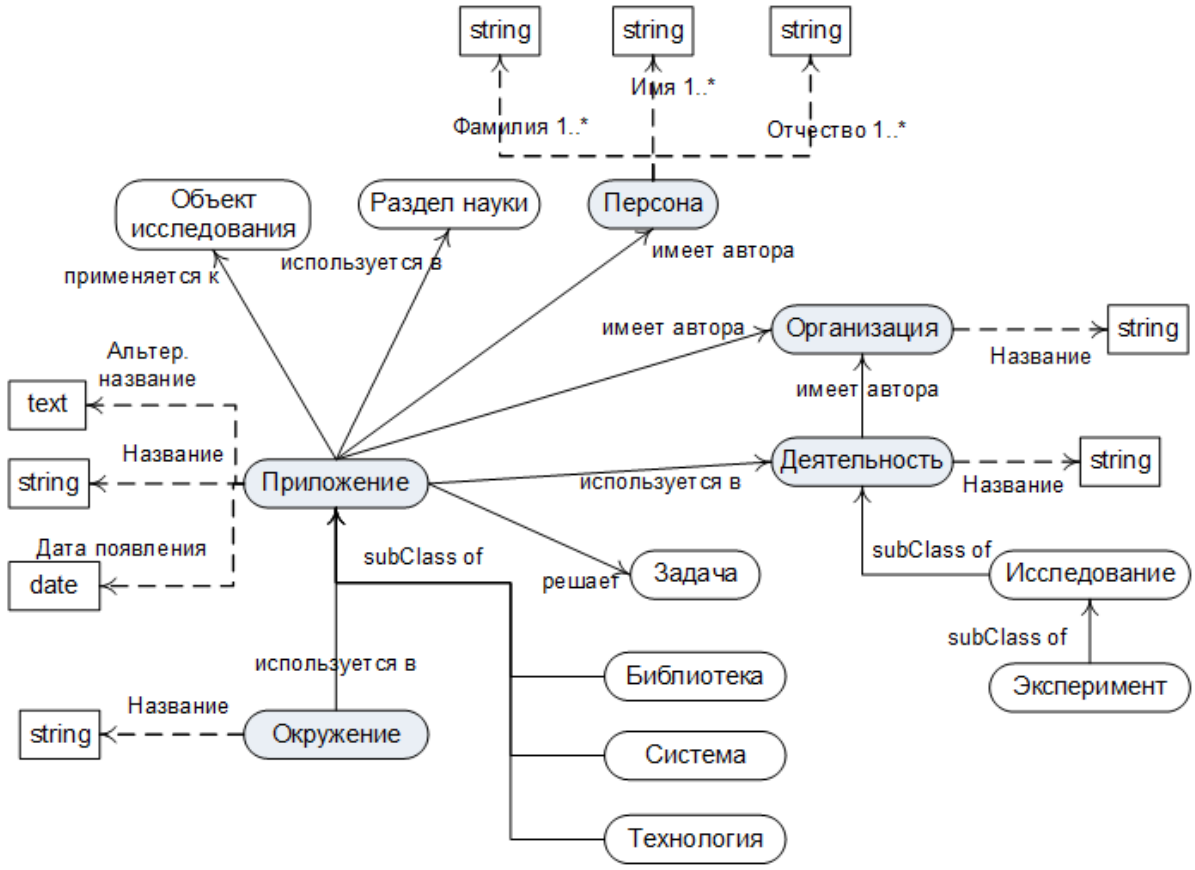

Рис. 3. Фрагмент онтологии, описывающий классы, связанные с научной деятельностью

G) В созданном наборе данных не оказалось представленным достаточного количества терминов и отношений, относящихся к классам *Информационный ресурс* (отдельно рассматривается его подкласс *Набор данных*), *Научный результат* и *Публикация*. Несмотря на это, было решено оставить данные классы в онтологии.

В итоге, созданная онтология по КЛ содержит 15 классов, 10 из которых уже содержались в базовой онтологии, и 111 отношений, 92 из которых присутствовали в онтологии до конкретизации. Для экспериментальных исследований классы, которые в текстах упоминались редко, были исключены из списка возможных категорий терминов.

4. Построение терминологического ядра онтологии

Построение терминологического ядра онтологии по компьютерной лингвистике включает в себя следующие этапы. Во-первых, необходимо составить систему классов предметного словаря, основанную на классах выбранной онтологии, и осуществить его начальное

наполнение базовыми терминами из онтологии и терминами из сопутствующих информационных ресурсов (тезаурусу и порталу по КЛ). На следующем этапе на основе составленного ранее корпуса текстов было необходимо разметить данные и создать датасеты для машинного обучения. На последующих этапах, можно непосредственно решать задачи извлечение и классификация терминов методами машинного обучения.

4.1. Генерация предметного словаря

Система классов в словаре генерируется на основе структуры онтологии, отражая иерархию ее объектов и отношений. Названия классов для терминов, обозначающих отношения, состоят из названий онтологических элементов в соответствии с шаблоном

<название_класса.название_отношения.название_класса>,

например: *Метод исследования.решает задачу.Задача*

Каждый термин словаря снабжается морфологической и семантической информацией, которые впоследствии используются при автоматической обработке текста.

Первоначальное наполнение предметного словаря велось в два этапа. На первом этапе был проанализирован русско-английский тезаурус по компьютерной лингвистике [26]. Из него были взяты подходящие термины классов: *Деятельность, Задача, Информационный ресурс, Метод исследования, Метрика, Объект исследования, Предмет исследования, Раздел науки* и *Результат*. Количество терминов составило 585.

Термины из тезауруса были сопоставлены с терминами на портале по компьютерной лингвистике [25], на котором расположена онтология научной деятельности, содержащая описанные выше классы. Терминам были добавлены подходящие онтологические классы, а тем, которых не оказалось на портале, были сопоставлены классы их синонимов или родовых понятий (более общее понятие для термина), если такие имелись. Оставшиеся термины были размечены вручную. Количество всех терминов, найденных на портале и в тезаурусе, составило 2640.

Для оценки качества автоматического сопоставления были просмотрены все термины на предмет неточного соотнесения с классом. Анализ показал, что неточности связаны с отсутствием класса в онтологии и неучтенной омонимии. Так, термин «*дискурс*» был отнесен к *Информационному ресурсу*, на основании существования *системы ДИСКУРС*. Это не является ошибкой, однако решено вручную добавить синонимичный термин и отнести его к классу *Деятельность*. Полнота автоматического соотнесения составила 75,9%, а точность – 99%. Высокая точность объясняется классификацией с опорой на синонимы и родовые

термины. При увеличении количества отношений в сопоставлении (не только синонимы и родовые понятия), полнота улучшалась, но точность сильно понижалась.

Поскольку тезаурус и портал не содержат терминов, появившихся в последние десять лет (в том числе относящихся к области машинного обучения), на следующем этапе необходимо было добавлять термины на основе корпуса современных текстов по КЛ.

4.2. Создание наборов данных

Для автоматического пополнения словаря терминами были составлены наборы данных на основе подобранных фрагментов текста из собранного корпуса по компьютерной лингвистике. Фрагменты текста подбирались так, чтобы получить репрезентативную выборку для всех классов терминов. На первом этапе для облегчения ручной разметки все тексты были автоматически размечены с помощью модели «ru_core_news_sm» из библиотеки Spacy. Использовалась BIO-разметка.

BIO-разметка текста (IOB-разметка) [8] – это способ разметки последовательных данных, таких как именованные сущности в тексте. Каждое слово в тексте помечается как "B" (начало сущности), "I" (продолжение сущности) или "O" (вне сущности). После пометки B, I или O через дефис указывается название класса. Так, для задачи извлечения терминов будет указываться B-TERM или I-TERM, а для задачи классификации для указания класса термина будут использоваться метки класса, например, к классу метод будут использоваться метки B-Method, I-Method.

Следующий этап заключался в отборе подходящих предложений, содержащих термины, относящиеся к нашей предметной области. Из них были составлены два набора данных: для задачи извлечения терминов и для задачи классификации терминов (размеченный по классам онтологии). Пример предложения в датасете для задачи извлечения терминов:

```
# sent_id = 744
# text = Создателем ORES является Wikimedia Foundation.
Создателем O
ORES B-TERM
является O
Wikimedia B-TERM
Foundation I-TERM
. O
```

Для второго набора данных для разметки терминов использовался список классов онтологии, описанный в предыдущих разделах, для которых метка имела вид B-Class для

начала термина, I-Class для остальных частей термина, а O указывало на часть вне термина, где Class - наименование онтологического класса. Пример предложения в датасете для задачи извлечения терминов:

```
# term_id = 5:
[TEXT]: <term>Извлечение именованных сущностей</term> из текста - одна из самых востребованных функций текстовой аналитики.
[TERM]: Извлечение именованных сущностей
[CLASS]: Task
```

Таким образом, в наших наборах было выделено 12 признаков для терминов классов, а объем каждого набора данных составил 1065 предложений (4180 терминов).

4.3. Извлечение терминов из корпуса

Для извлечения терминов из текста и отнесения их к соответствующим классам был применен нейросетевой подход на основе архитектуры трансформер. Задача извлечения терминов из текста относится к классу NER (Named Entity Recognition): каждое отдельное слово может быть либо началом термина, либо продолжением термина, либо не термином.

В качестве нейросетевой модели была рассмотрена XLM-RoBERTa, предобученная на 2.5 терабайтах данных из корпуса CommonCrawl на 100 различных языках. Как видно из названия, данная модель относится к семейству моделей RoBERTa, впервые она была представлена в [1]. В дополнение к слоям трансформера были добавлены линейный слой и softmax для осуществления классификации токенов (Рис. 4).

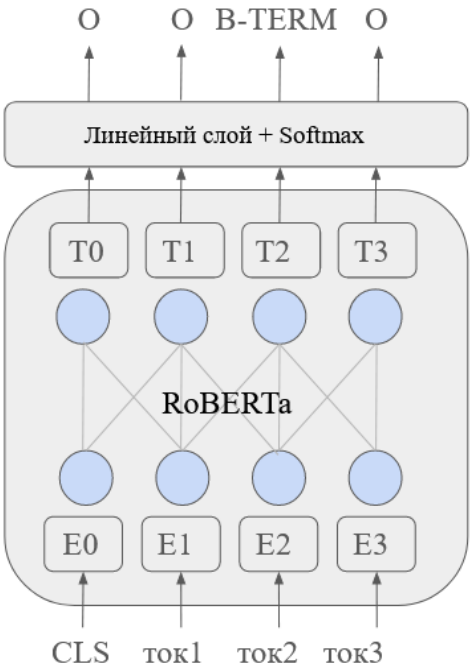

Рис. 4: Структура классификатора для извлечения терминов

Отдельно стоит отметить, что в общем случае токенизатор преобразует слово в несколько токенов, поэтому каждому токену присваивается метка класса исходного слова. Затем, когда классификатор установит метку каждого токена, метка слова определится как самая частая метка среди его токенов.

Помимо модели xlm-roberta-base в исследовательских целях рассматривались также следующие модели: roberta-base-russian-v0, предобученная на корпусе русскоязычных текстов «Тайга» [12], и ruRoberta-large, представленная среди семейства русскоязычных трансформеров в [18].

Обучающая и тестовая выборки получились путем разделения исходного набора данных в соотношении 9 к 1. Полученные результаты описаны в Таблице 1. Использование модели ruRoberta-large позволяет достичь показателя F1-меры в 91%.

Таблица 1. Значения метрик для задачи извлечения терминов

Модель	Полнота	Точность	F1-мера
roberta-base-russian-v0	0.81	0.76	0.80
xlm-roberta-base	0.86	0.83	0.85
ruRoberta-large	0.92	0.90	**0.91**

Была проведена классификация выявленных ошибок, допускаемых моделью. Примеры ошибок и их процент от совокупности всех ошибок представлены в Таблице 2. Несмотря на

то, что выделение середины термина является частным случаем ошибок, возникающих при извлечении вложенных сущностей, мы выделили ее отдельно. Это объясняется тем, что именно эта разновидность нуждается в корректировке и постобработке.

Таблица 2. Примеры ошибок разметки терминов моделью

Тип ошибки	% ошибок	Термин	Ожидаемое значение	Полученное значение
Не извлечено	41%	*классификация*	B-TERM	O
Новый термин	26,8 % (21,4%)	*модель сентиментного анализа*	O O O	B-TERM I-TERM I-TERM
Вложенная сущность	24,1%	*метод морфологического анализа*	B-TERM I-TERM I-TERM	O B-TERM I-TERM
Выделение середины термина	6,3%	*логистическая регрессия*	B-TERM I-TERM	O I-TERM
Выделение однородных членов	0,9%	*лексические и синтаксические признаки*	B-TERM O B-TERM I-TERM	B-TERM O B-TERM O
Пропуск середины термина	0,9%	*программу машинного перевода*	B-TERM I-TERM I-TERM	B-TERM O I-TERM

Анализ ошибок показал, что в 45,4%, несмотря на неточное извлечение термина моделью, эксперты отметили их как допустимые. К таким случаям относятся выделение вложенных сущностей и некоторых новых терминов (эксперты отметили их как допустимые в 21,4%случаев).

Для решения проблемы выделения середины термина в рамках постобработки возможно изменение метки I-TERM на B-TERM. Это мотивировано тем, что выделяется не вся сущность, а вложенная в нее другая сущность. Так, например, возможно выделения термина *регрессия*, поскольку *логистическая регрессия* является ее частным случаем.

Пример извлечения однородных групп “*лексические и синтаксические признаки*” показывает необходимость их постобработки и извлечения нескольких словосочетаний с общим словом(-ами): *лексические признаки* и *синтаксические признаки.*

Решение проблемы пропуска середины термина возможно объединением во время постобработки всего словосочетания от ближайшего начала (B-TERM) до последнего последовательного идущего указания на часть термина (I-TERM).

4.4. Классификация терминов

Следующим шагом была классификация извлеченных терминов. В качестве данных использовалась исходная выборка, однако вместо общих меток для терминов использовались метки конкретных классов.

Классифицировать термин в отрыве от контекста было бы некорректно, в связи с этим было принято решение передавать модели все предложение, в котором термин будет выделен специальным образом — это позволит модели сфокусировать основное внимание именно на нем. Таким образом, классифицироваться будет весь текст, который передается на вход модели. Структура классификатора представлена на Рис. 5.

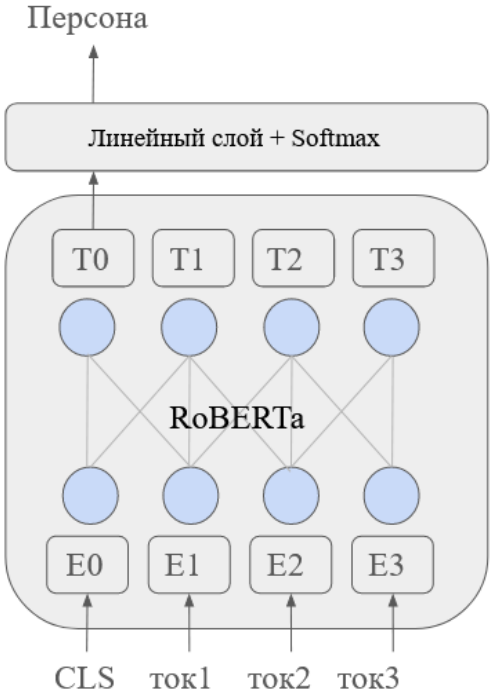

Рис. 5: Структура классификатора терминов

Для упрощения классификации некоторые похожие классы были объединены в один, так, например, классы *Приложение*, *Библиотека*, *Технология*, *Окружение* были объединены в один общий класс *R-Приложение*, поскольку они все так или иначе имеют семантику приложения. Аналогично *Набор данных* и *Корпус* были объединены в *R-Набор данных*. Итоговый список классов: *Метод, Деятельность, Объект, Персона, Задача, Организация, Модель, Метрика, Значение, Дата, Язык, R-Набор данных, R-Приложение, Раздел науки.*

В качестве модели для классификации была выбрана ruRoberta-large, показавшая лучшие результаты для задачи извлечения терминов. В результате среднее значение F1-меры составило 89%. Для каждого отдельного класса значение представлено на Рис. 6.

Для четырех классов F1-мера составила 100%, еще для шести F1-мера превысила 85%. Однако распознавание объектов класса *Деятельность* модели дается с трудом: значение составляет лишь 66%.

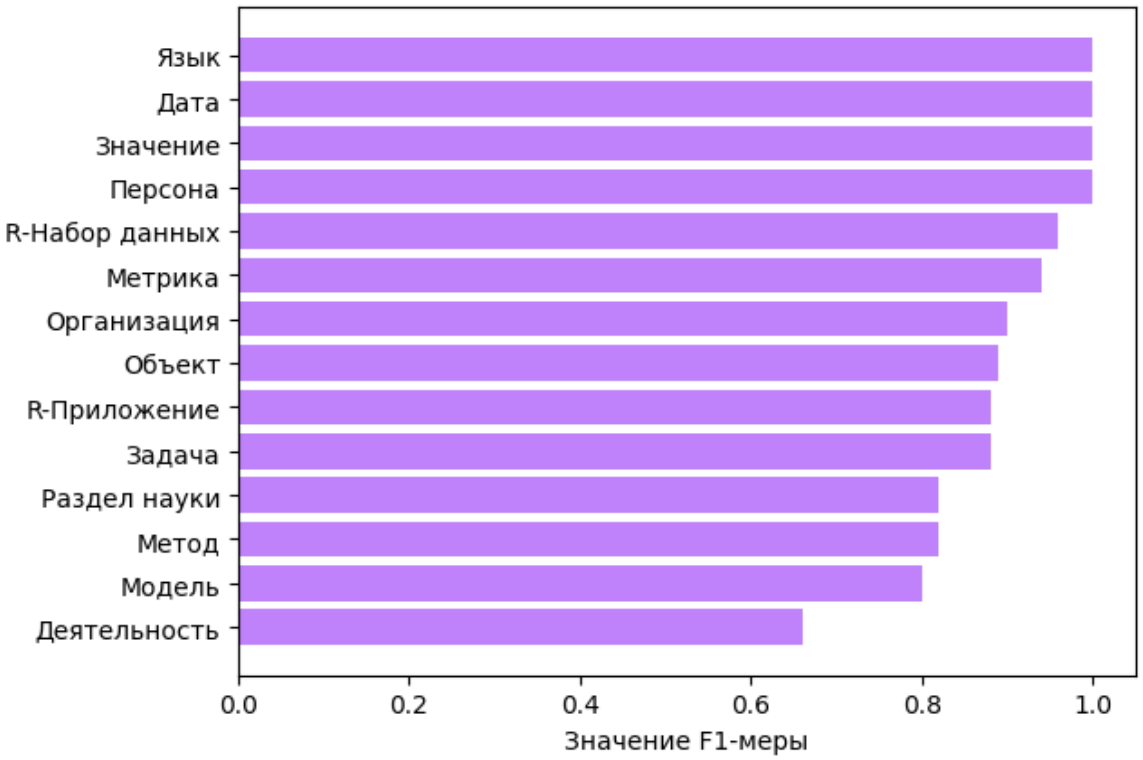

Рис. 6: Значения F1-меры для классов терминов

Анализ ошибок показал сложности для модели в различении названий организаций и названий приложений. Так, например, термин *Яндекс.Браузер* был определен как название организации, а не приложения, а термин *Майкрософт* модель напротив выделила как название приложения. Также иногда возникают сложности с тем, чтобы различать термины, которые относятся к классам Задача, Раздел науки и Деятельность. Так, например, в отрывке из текста "*Для задачи <term>моделирования</term> языка ULMFit использует <...>*" модель отнесла *моделирование* к классу Деятельность, а не классу Задача.

5. Архитектура системы пополнения онтологии

Переход от традиционных способов разработки онтологий, когда основным источникам знаний выступает эксперт, к интеллектуальным технологиям, обеспечивающих извлечение

знаний из неструктурированных источников данных, привел к возникновению технологий, в которых роль эксперта заключается в решении совсем других задач: проектировании концептуальных верхнеуровневых абстракций, разметке данных (для использования методов машинного обучения) и валидации полученных результатов. Процесс формирования онтологий на основе неструктурированного контента получил название обучение онтологий (ontology learning). Разрабатываемая система для обучения и дальнейшего пополнения онтологии включает ряд задач (Рис. 7).

Рис. 7: Основные этапы для решения задачи пополнения онтологии

На первом этапе необходимо определиться с онтологией верхнего уровня. Можно взять уже существующую [15], либо использовать паттерны онтологического проектирования для описания основных классов онтологий [2] научной области знаний: *Метод исследования, Задача, Модель обучения, Набор данных, Метрика, Приложение, Окружение, Персона,*

Научный результат, Раздел науки, Объект исследования, Предмет исследования, Публикация, Деятельность, Информационный ресурс.

Второй этап заключается в создании корпуса текстов по выбранной предметной области. На его основе проходит конкретизация онтологии на выбранную область: обобщаются и формируются новые классы. Также фиксируются фрагменты текстов, демонстрирующие примеры употребления для новых классов, отсутствующих в базовой онтологии, для которых впоследствии описываются паттерны содержания.

Следующий этап заключается в создании предметного словаря. На основе онтологических классов, атрибутов и отношений генерируется система лексико-семантических классов словаря. Для автоматического пополнения словаря предлагается использовать модели машинного обучения. В соответствии с этим необходимо создание набора данных на основе фрагмента собранного корпуса текстов. Извлеченные термины классифицируются в соответствии с лексико-семантическими классами и добавляются в словарь. Для пополнения словаря также можно использовать ресурсы близкой тематики (WikiData, тезаурусы, порталы и т.д.). Они сопоставляются друг с другом, после чего уникальные термины добавляются в словарь и снабжаются подходящими онтологическими классами.

Следующие этапы не рассматривались в рамках данной работы. На пятом этапе происходит генерация вопросов оценки компетентности (ВОК). Они могут быть использованы для извлечения отношений из текстов с использованием моделей машинного обучения, либо для извлечения грамматических ограничений для лексико-синтаксических паттернов [24].

На следующем этапе происходит генерация лексико-синтаксических паттернов для извлечения новых терминов и объектов предметной области из текста. Генерация может происходить в том числе с помощью мета-паттернов и онтологической информации [22].

На последнем этапе сгенерированные шаблоны, словарь и корпус используются для пополнения онтологии.

Заключение

Данная работа проводилась в рамках общего проекта по созданию автоматизированных методов построения онтологий [22, 14]. Особенностью рассматриваемого подхода является использование базовой онтологии, конкретизируемой на предметную область, и предметного словаря, а также пополнение словаря с использованием методов, основанных на глубоком обучении. Подход применялся для построения онтологии современных методов компьютерной лингвистики.

Применение корпусных методов анализа понятий базовой онтологии научной области знаний позволил выявить несоответствия и провести конкретизацию онтологии, в частности добавлены новые классы, связанные с методами машинного обучения. Для доказательства обоснованности изменений использовался подсчет встречаемости терминов в корпусе текстов. В результате данной работы создана онтология по КЛ, которая включает в себя 15 классов и 111 отношений. Онтология представлена в формате OWL и будет выложена в открытый доступ.

Другой особенностью является подход к созданию и пополнению предметного словаря. Система классов генерируется на основе структуры онтологии в соответствии со специализированным шаблоном. Для пополнения словаря были сопоставлены русско-английский тезаурус и портал по компьютерной лингвистике. Из их объединения были взяты уникальные термины, которым впоследствии были выведены классы. Точность и полнота автоматического соотнесения 99% и 75,9% соответственно. Количество терминов в словаре составило 2640.

В целях пополнения терминологического ядра онтологии был применен нейросетевой подход. С этой целью был создан размечен набор данных (датасет), включающий 1000 предложений из собранного корпуса текстов по компьютерной лингвистике с BIO-разметкой[1]. Рассматривались две подзадачи: извлечение терминов и их классификация. В задаче обнаружения терминов получилось достичь уровня 91% F1-меры на тестовой выборке. Анализ ошибок показал, что часто модель выделяет термин только частично, поэтому целесообразно будет проводить дальнейшие исследования с учетом этой особенности (например, провести постобработку на основе правил). Задача классификации терминов на 12 классов была осложнена небольшим размером набора данных для каждого класса, что является следствием достаточно трудоемкого процесса ручной разметки текстов. Тем не менее средний показатель F1-меры на тестовой выборке составил в среднем 89%, что является довольно неплохим результатом для начального исследования, однако в дальнейшем планируется рассмотреть различные способы улучшения качества результатов и для данного этапа.

Создаваемая терминологическая система является основой для дальнейшего построения и пополнения онтологии по компьютерной лингвистики, качество которой в значительной степени зависит от корректности выделения терминов. В качестве развития темы планируется также рассмотреть возможность извлечения названий отношений и значений атрибутов.

[1]Наборы данных доступен по ссылке: https://github.com/pasukka/NLP-Dataset.git

Список литературы

1. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale, 2019. doi: https://doi.org/10.48550/arXiv.1911.02116.
2. Association for Ontology Design & Patterns. Режим доступа: http://ontologydesignpatterns.org (дата обращения: 18.09.2018).
3. Blomqvist E., Hammar K., Presutti V. Engineering Ontologies with Patterns: The eXtreme Design Methodology // Ontology Engineering with Ontology Design Patterns. Studies on the Semantic Web. IOS Press, 2016. P.23-50.
4. Fernández-López M., Gómez-Pérez A., Pazos A., Pazos J. Building a Chemical Ontology Using Methontology and the Ontology Design Environment. IEEE Intelligent Systems & their applications, 1999, 4(1). P. 37–46.
5. Gangemi A., Presutti V. Ontology Design Patterns // Handbookon Ontologies. Springer, 2009. P. 221-243.
6. Hagen Soltau, Izhak Shafran, Mingqiu Wang, Laurent El Shafey. RNN Transducers for Nested Named Entity Recognition with constraints on alignment for long sequences, 2022. doi: https://doi.org/10.48550/arXiv.2203.03543.
7. Juliao Braga, Joaquim L. R. Dias, Francisco Regateiro. A Machine Learning Ontology, 2023. doi: https://doi.org/10.31226/osf.io/rc954.
8. Lance A. Ramshaw, Mitchell P. Marcus. Text Chunking using Transformation-Based Learning, 1995. P. 82-94. arXiv:cmp-lg/9505040
9. Li J., Sun A., Han J., Li C. “A survey on deep learning for named entity recognition”, IEEE Transactions on Knowledge & Data Engineering. , 2022. Vol. 34, no. 1. P. 50-70.
10. Maynard D., Funk A., Peters W. Using lexico-syntactic ontology design patterns for ontology creation and population. In Proc. of WOP2009 collocated with ISWC2009, V. 516. pp. 39-52, CEUR-WS.org.
11. Piskorski J., Babych B., Kancheva Z., Kanishcheva O., Lebedeva M., Marcinczuk M., Nakov P., Osenova P., Pivovarova L., Pollak S., et al., “Slav-ner: The 3rd cross-lingual challenge on recognition, normalization, classification, and linking of named entities across slavic languages”, in Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing, 2021. P. 122-133.
12. Shavrina T., Shapovalova O. (2017) To the methodology of corpus construction for machine learning: «taiga» syntax tree corpus and parser. in proc. of “CORPORA2017”, international conference , Saint-Petersbourg, 2017.
13. Sure Y., Staab S., Studer R. On-To-Knowledge Methodology // Handbook on Ontologies. 2003. № 6 P.135–152.

14. Uschold M., King M. Towards a Methodology for Building Ontologies // Proceeding of the Workshop on Basic Ontological Issues in Knowledge Sharing. Montreal, Canada. 1995. P. 6.1–6.10.
15. Zagorulko Y. A. Using a System of Heterogeneous Ontology Design Patterns to Develop Ontologies of Scientific Subject Domains // Programming and Computer Software, 2020. P. 273-280.
16. Zhu F, Shen B. Combined SVM-CRFs for biological named entity recognition with maximal bidirectional squeezing, 2012. doi: https://doi.org/10.1371/journal.pone.0039230.
17. Zihan Liu, Feijun Jiang, Yuxiang Hu, Chen Shi, Pascale Fung. A Pre-trained Model for Low-Resource Entity Tagging, 2021. doi: https://doi.org/10.48550/arXiv.2112.00405.
18. Zmitrovich D., Abramov A., Kalmykov A., Tikhonova M., Taktasheva E., Astafurov D., Baushenko M., Snegirev A., Shavrina T., Markov S., Mikhailov V., Fenogenova A. A Family of Pretrained Transformer Language Models for Russian, 2023. doi: https://doi.org/10.48550/arXiv.2309.10931.
19. Большакова Е.И., Баева Н.В., Бордаченкова Е.А., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны в задачах автоматической обработки текстов. Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2007. М.: Издательский центр РГГУ, 2007. С. 70-75.
20. Загорулько Ю. А., Боровикова О. И., Кононенко И. С., Соколова Е. Г. Методологические аспекты разработки электронного русско-английского тезауруса по компьютерной лингвистике // Информатика и её применения. 2012. № 6(3). С. 22-31.
21. Загорулько Ю.А. Построение порталов научных знаний на основе онтологий // Вычислительные технологии, спецвыпуск 2. 2007. Т. 12.
22. Кононенко И.С., Сидорова Е.А. Методика разработки лексико-семантических паттернов для извлечения // Системная информатика. 2022. № 20. С. 25-46.
23. Лагутина Н.С., Васильев А.М., Зафиевский Д.Д. Задачи в области распознавания именованных сущностей: технологии и инструменты. Моделирование и анализ информационных систем. 2023. № 30(1). С. 64-85.
24. Овчинникова К. А. Автоматическая генерация лексико-синтаксических паттернов на основе онтологии для извлечения информации о научной деятельности // Литературоведение. Прикладная лингвистика. Языкознание: Материалы 60-й Междунар. науч. студ. конф. Новосибирск: ИПЦ НГУБ, 2022. С. 203-205.
25. Портал по компьютерной лингвистике. Режим доступа: https://uniserv.iis.nsk.su/cl (дата обращения: 07.04.23)
26. Русско-английский тезаурус по компьютерной лингвистике. Access mode: https://uniserv.iis.nsk.su/thes (дата обращения: 07.04.23)
27. Соловьев В. Д., Добров Б. В., Иванов В. В., Лукашевич Н. В. Онтологии и тезаурусы (учебное пособие). Казань, Москва, 2006. 157 с.