УДК 004.89 DOI: 10.17587/it.30.622-632

И. Ю. Каширин, д-р техн. наук, проф. e-mail: igor-kashirin@mail.ru,
Рязанский государственный радиотехнический университет имени В. Ф. Уткина

Токенизация политических текстов в BERT-моделях с использованием ICF^{+}-онтологий

Рассматривается проектирование языковых моделей машинного обучения, а также их ансамблей, применяемых в сложной аналитике новостных текстов отечественных и западных электронных средств массовой информации. Приводится пример программной реализации новой языковой нейросетевой модели с проблемно-ориентированной онтологической токенизацией. В качестве инструментария используется язык Python v.3.10, Anaconda v.2.1. Эффективность подхода в сравнении с лучшими зарубежными аналогами подтверждается серией экспериментов на примере классификации новостных статей по их идеологической направленности на западные и англоязычные российские.

Ключевые слова: *Bert-модели, онтологические модели, ICF^{+}-отношение, токенайзер, ретривер, политические новости, ансамбли ML-моделей, прогнозирование, семантическое сходство*

Введение

Информационная война, которая в течение десятилетий ведется проамериканскими средствами массовой информации (СМИ), уже привела к фактической потере суверенитета странами Западной Европы, Японией, Австралией и многими другими.

Более 90 % всех мировых СМИ, включая Web-ресурсы, газеты, журналы, телевидение, принадлежат следующим крупным корпорациям: Time Warner, News Corporation, The Walt Disney Company, Viacom/CBS Corporation, Comcast/NBCUniversal, полностью отражающим политическую идеологию США. Идеологическая обработка общества с использованием информационных технологий является весьма высокотехнологической и наукоемкой сферой деятельности, использующей особенности человеческой психологии и основы теории больших данных. Фейк-новости или подача материала в идеологизированной форме стали закономерным предметом исследований ученых, работающих в сфере Data Mining с нейронными сетями и другими моделями машинного обучения (ML-моделями).

Широкое применение в автоматизации интеллектуального исследования электронных публикаций нашли BERT-модели (Bidirectional Encoder Representations from Transformers), основанные на нейронных сетях и спроектированные компанией Google AI в 2018 г. Главным принципом этих нейросетевых моделей является использование технологии трансформеров как моделей для анализа зависимостей в сложных естественно-языковых конструкциях, таких как части предложений и даже предложения полностью.

Архитектурный принцип разбиения задач на подзадачи в модели BERT применяется в форме решения двух подзадач генерации текстов:

— заполнение в текстах пропущенных фрагментов на уровне слогов, слов и словосочетаний;

— предсказание предложения, следующего за определенным фрагментом текста.

Эффективное решение этих задач возможно только при использовании текстовых корпусов, насчитывающих миллионы предложений и больше.

Исследование, представленное в настоящей статье, посвящено созданию новой технологии мониторинга англоязычных СМИ. На текущий момент времени достижения в области проектирования IT-средств интеллектуального анализа естественно-языковых текстов позволяет использовать BERT-модели в задачах классификации и регрессии статей СМИ по

перечисленным ниже классам психологической окраски текстов:

— достоверные или сфальсифицированные новости [1];

— материалы токсичного характера [2];

— тексты с сарказмом и иронией [3];

— гневные статьи [3, 4];

— тексты, способствующие социальному взрыву [4];

— статьи с отрицательной или положительной эмоциональной окраской [5];

— тексты с комбинированным многоаспектным содержанием [6].

Предобучение моделей дает возможность получить готовую модель, которую можно использовать в вопросно-ответных системах или системах классификации для заранее выбранной предметной области. Для моделей высокой мощности, таких как модель OpenAI с триллионом параметров [7], предметная область весьма широка, но все равно остается ограниченной. То же самое относится к российской генеративной языковой модели Яндекса YaLM и языковой модели ruGPT3XL.

Большие модели обучаются на огромных множествах текстов, среди которых оказываются и тексты, содержащие неточные, неверные и часто противоречивые знания. Вследствие сказанного выбор данных для обучения является отдельной и довольно непростой задачей, состоящей из двух этапов:

— предобучение на текстовом корпусе больших данных (pre-training);

— дообучение на сравнительно небольшом множестве проблемно-ориентированных текстов (fine tuning).

Такие модели становятся не только эффективным подспорьем при разработке новых высокотехнологичных проектов, но и оружием информационной войны. Более 86 % материалов СМИ (текстовых корпусов, впоследствии используемых для обучения больших нейросетевых моделей) носят выраженную прозападную окраску. Это же относится к международным репозиториям данных [8, 9]. Актуальным вопросом является определение возможности решения перечисленных задач с помощью относительно дешевых вычислительных мощностей на основе подмножеств текстовых корпусов умеренных размеров.

Интегрирование рассмотренных проблем на основе единого программного инструментария, основу которого составляют BERT-модели, позволяет обобщить *цель разработки* новой технологии как создание средств интеллектуального мониторинга материалов СМИ.

Для достижения этой цели предлагается использовать механизм проблемно-ориентированной токенизации текстов, базирующейся на специфических свойствах онтологий наиболее популярных тематических направлений в информационно-политической области. В настоящей статье выделяется задача классификации материалов современных СМИ по идеологическому основанию на западные (идея однополярного мира) и восточные (интернациональная идея).

1. Мониторинг электронных СМИ

В основе мониторинга СМИ лежит применение рекуррентных двунаправленных нейронных сетей и BERT-моделей с их дообучением с помощью онтологических моделей знаний [10]. Проблемная ориентация анализа при мониторинге СМИ должна опираться на родовидовые, причинно-следственные таксономии, таксономии типа "часть—целое" и другие, исследованные в рамках дескриптивных логик [11, 12]. Практическая реализация такого подхода возможна с помощью языков разметки, таких как OWL или RDF. Немало способствуют внедрению языков разметки и графические средства Protégé 4.0.2, которые используются автором при дальнейшем изложении.

Международными научными школами, в частности Стенфордским университетом [13], были разработаны онтологии в соответствующих пространствах имен. Эти пространства содержат семантические описания большого числа понятий, отношений и процессов для различных предметных областей. Сложностью применения OWL и RDF следует считать существование многочисленных описаний одних и тех же предметных областей различными авторскими коллективами, в результате чего описания неизбежно становятся взаимно противоречивыми. Одни и те же концепты могут рассматриваться в разных пространствах имен на противоположных местах одного и того же отношения, например, Is-A(attack, conflict) "атака является конфликтом" или Is-A(conflict, attack) "конфликт является атакой". Кроме того, семантика дескриптивной логики при всей своей сложности является предикативной и не позволяет использовать полиморфическое представление знаний [14]. Эти факты свидетельствуют о необходимости проектирования новых онтологий, допускающих полиморфическое представления знаний с разных углов

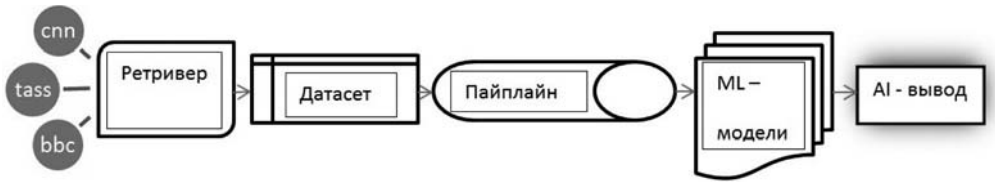

Рис. 1. Общая схема мониторинга электронных СМИ

зрения, но непротиворечиво локализующих семантику одного текста.

В новой технологии предлагается применять программный инструментарий WordNet и Word2vec, а также пакет программ spaCy, которые позволяют максимально снизить трудоемкость проектирования онтологий и их последующее использование в BERT-моделях.

Общая схема мониторинга электронных ресурсов изображена на рис. 1.

Схема отображает последовательность действий при мониторинге изданий, которая начинается с создания корпуса текстов на основе применения ретривера CorpusMining v.2.1, программно реализованного автором статьи. Ретривер является интеллектуальным поисковым агентом и собирает статьи разного времени и различных тематик по нескольким темам политической направленности. Полученные датасеты проходят этап подготовки данных, для чего используется соответствующий пайплайн, включающий токенизацию, лемматизацию и NER-обработку предложений каждой политической публикации. Для обеспечения сбалансированности текстового корпуса исследованию подлежат как зарубежные, так и российские англоязычные СМИ.

Идентификация материалов СМИ по классам идеологической и психологической окраски текстов реализуется модернизированными BERT-моделями. На заключительном этапе мониторинга оценивается существующая политическая ситуация и прогнозируется дальнейшая динамика политических событий.

2. Web-скрапинг

Web-скрапинг — это процедура извлечения данных, релевантных целевому запросу, использующая функционал ретривера CorpusMining v.2.1, а также средства среды Python v.3.6: BeautifulSoup4 v.4.12.3 и Googlesearch v.3.0.

Для проведения серии экспериментов с обучением нейронной сети и последующей классификации ретривером были сформированы корпуса текстов общим числом более 5000 электронных статей. Автоматический поиск на основе ключевых слов использовал западные издания: Aljazeera, Bloomberg, cnn, nytimes, WSJ, theguardian. Из российских англоязычных издательств были выбраны: en.kremlin, medusa, RT, tass, Interfax.

3. Создание тематической онтологии "террористический акт"

Методология разработки ретроспективных и тематических моделей данных в форме owl-онтологий является новой и может быть использована разработчиками для повышения точности моделей машинного обучения в предметных областях, аналогичных области политических новостей. В частности, таким новшеством можно считать способ проблемно-ориентированной токенизации, рассмотренный далее.

Токенизация как первоначальный этап обработки текста включает в себя векторизацию естественно-языковых предложений и специализированную разметку лексических конструкций. Векторизация решает задачу замены слов предложения на числовые значения. При этом наиболее часто упоминаются следующие методы:

— мешок слов (bag of words);

— TF-IDF (Term Frequency, частотность слова в документе, и Inverse Document Frequency, инверсия частоты документа);

— встраивание слов (word embeddings).

Наиболее эффективным показал себя последний метод встраивания слов, поскольку в его основе лежит весьма близкое к моделям представления знаний понятие семантического пространства. В n-мерном семантическом пространства близкие по смыслу слова располагаются рядом.

В рассматриваемой здесь технологии используется новый метод встраивания слов, основанный на теории иерархических чисел, впервые введенной в 2020 г. [15] для разметки концептов модели знаний в родовидовых таксономиях [16]. Оказалось, что иерархические числа, нормализованные до промежутка [0, 1], при грамотном построении таксономий могут не уступать, а в некоторых экспериментах и

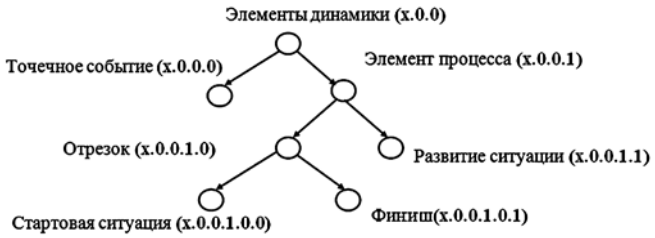

Рис. 2. Пример разметки событийной таксономии иерархическими числами

превосходить существующие методы встраивания слов.

Краткий пример разметки событийной таксономии приведен на рис. 2.

Здесь вершины дихотомической таксономии помечены бинарными иерархическими числами, для каждой дочерней вершины используется индекс отцовской, дополненный цифрами 0 (левый потомок) или 1 (правый потомок).

При использовании предобученных моделей глубокого обучения их эффективность можно повысить на этапе токенизации текста, использовав вставку спецтокенов предметной области.

В качестве предметной области рассмотрим политические новостные тексты, онтологии для которых содержат три основных таксономии: родовидовую (гипонимы/гиперонимы), причинно-следственную (каузуальную) и стратификационную (меронимы, часть—целое). Рис. 3 демонстрирует пример графического представления онтологии "террористическая атака", построенного в редакторе Protege 4. Здесь введены два главных отношения "people (civilian, military)" ("люди (гражданские, военные)"), которое может интерпретироваться следующими текстовыми выражениями:

— люди могут быть гражданскими лицами;

— люди могут быть военными;

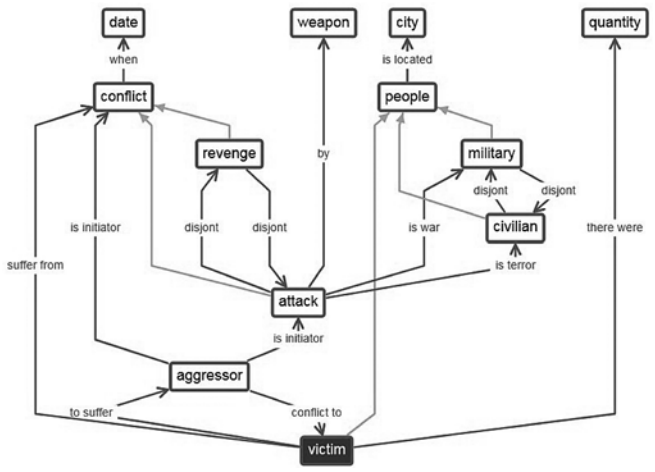

Рис. 3. Онтологическая схема для предметной области "террористическая атака"

— гражданские и военные — это разные люди;

— гражданские лица как любые люди могут становиться военными;

— военные как любые люди могут становиться гражданскими.

То же самое касается отношения "conflict (revenge, attack)" ("конфликт(месть, нападение)"), т. е. "нападение может быть местью, а месть — нападением".

Такое отношение с множественным смыслом является представителем ICF $^{+}$-отношений [10]. Его можно задать с помощью следующего выражения:

$$ICF^{+}(X, y, z) = IsA(X, y) \cap IsA(X, z) \cap$$
$$\cap Cont(y, z) \cap Form(X, y) \cap Form(X, z).$$

Составляющими производного трехместного отношения ICF^{+} [**I**sA,**C**ont,**F**orm] являются базовые бинарные отношения:

— IsA — родовидовое отношение;

— Cont — отношение противопоставления концептов (в точной семантике не совпадающее с disjoint в языке OWL);

— Form — отношение "проявляться в форме".

Замечательное отношение Form семантически интерпретирует тройку (X, y, z) как взаимное наследование концептами y и z всех свойств друг друга опосредовано через концепта-предка X. Действительно, прозападными СМИ почти все атаки США трактуются как месть за какую-либо провинность будущей жертвы.

Вторичные отношения онтологической схемы соответствуют отношениям следующей интерпретации:

— результатом конфликта являются жертвы;

— в конфликте используется оружие;

— начало конфликта идентифицируется датой.

Посмотрим далее, как приведенный фрагмент онтологии будет использован в проблемно-ориентированной токенизации новостных политических текстов, например, следующего фрагмента, процитированного из издания "CNN":

"March 23, 2024 Shooting at Moscow concert venue leaves over 130 dead."

("23 марта 2024 в результате стрельбы на концертной площадке в Москве погибло более 130 человек.")

В соответствии с рассматриваемой технологией мониторинга информационных ресурсов для анализа приведенного текста используется библиотека spaCy с импортом nlp-парсера и построителя теоретико-графовых отношений networkx:

```
import spacy
sentences = getSentences(text)
nlp_model = spacy.load('en_core_web_sm')
triples = []
for sentence in sentences:
  triples.append(processSentence(sentence))
printGraph(triples)
```

Результатом их применения на втором этапе технологии будет множество графов, которые в сокращении представлены фрагментом на рис. 4.

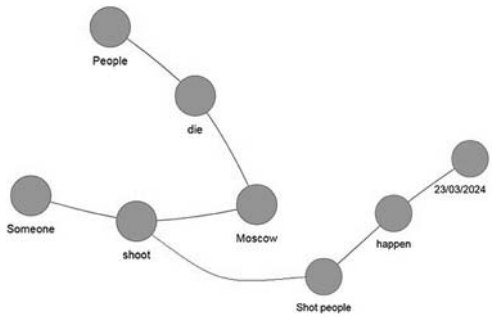

Рис. 4. Фрагмент прецедента онтологии "акт агрессии" для новости из СМИ "CNN"

Приведенный в примере граф соответствует прецеденту OWL-онтологии (см. рис. 3). Для формального установления этого соответствия используются алгоритмы унификации, рассматривающие каждый концепт как переменную с областью определения во множестве синонимов и гипонимов этого концепта.

Приведем некоторые области определения:

Dom(city) = { Moscow, Ryazan, Palestine, Kiev, London, ...};

Dom(attack) = { shooting, bombing, blockade, arson, ... };

Dom(quantity) = N (1, 2, 3, ...);

Dom(victim) = { injured, killed, dead, ... }.

В синтаксисе OWL можно задать не только области определения концептов как OWL-классов, но и выразить определение террора как атаку со многими гражданскими жертвами:

```
<owl:ObjectProperty rdf:ID = "isTerror">
  <rdfs:domain rdf:resource = "#Civilian"/>
  <rdfs:range rdf:resource = "#People"/>
</owl:ObjectProperty>
<owl:Class rdf:about = "#Victim">
  <rdfs:subClassOf rdf:resource = "&owl;Thing"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource = "#there Were"/>
      <owl:cardinality rdf:datatype =
      "&xsd;nonNegativeInteger">
        1000000 </owl:cardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

Для унификации в новой технологии используется механизм паттернов (pattern) библиотеки spaCy.matcher:

Например, если унифицирован паттерн:

```
[{"OP":"*"}, {"LEMMA": "attack"}{"OP":"*"},
{"ENT_TYPE": "GPE"},
{"OP":"*"},{"LEMMA": "damage"}],
```

это означает, что вхождение в текст леммы GPE — это географический объект ("Moscow"), являющийся местом атаки, а присутствие ключевых слов из списков ["shooting", "bombing", "blockade", "arson"] и ["damaged", "died", "destroyed"] позволяет ассоциировать их с классами "attack" (нападение) и "damage" (повреждение). При включении этих гипонимов в онтологию сформируется новый фрагмент, приведенный на рис. 5.

Эта онтологическая модель позволяет упростить описание основных классов и отношений текста и дает возможность выделения в статьях отдельных фактов. Таким образом, появляется возможность проводить исследования, связанные с достоверностью и токсичностью материалов различных изданий. Эта онтология существенно проще OWL-описания общей онтологии, изображенной на рис. 3.

Одним из основных средств, позволяющих повысить точность и достоверность BERT-моделей, является модернизация этих моделей в части модуля проблемно-ориентированной токенизации.

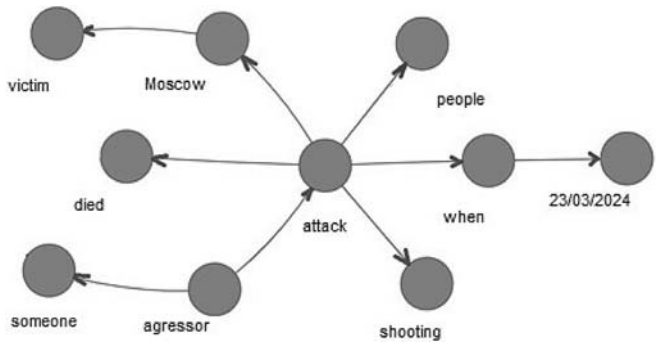

Рис. 5. Фрагмент онтологии с гипонимами

4. Проектирование шаблонов для проблемно-ориентированной токенизации

Предметная область классификации политических новостей весьма обширна, однако можно выделить несколько наиболее ярких событий, для которых можно разработать прикладную онтологию, *ограничивающую пространство поиска материалов* и соответствующий словарь политических терминов. Такими событиями могут быть, например, следующие:

— встреча государственных политических деятелей;

— начало военного конфликта;

— террористическая атака;

— встреча блокового объединения государств.

Для каждой такой темы можно получить ограниченный словарь, который можно разделить по семантическим словарным категориям. Каждая такая категория должна получить уникальное наименование или метку, которая послужила бы спецтокеном при токенизации предложений из материалов СМИ.

Таким образом, задача модернизации моделей машинного обучения становится более узкой и сводится к вставке в анализируемые тексты специальных токенов перед словарными конструкциями для выделения их значимости при классификации текстов. Можно с уверенностью предположить, что это приведет к сокращению времени классификации и повышению точности моделей. Вставка реализуется применением шаблонов (паттернов) библиотеки spacy.matcher до начала работы токенайзера нейронной сети.

Для вставки спецтокенов используется токенайзер предобученной модели bert-base-uncased. Модель, представляемая файлом pytorch_model.bin, применяется для точной настройки позже. Однако для определения семантического сходства понятий, используемых в новостных публикациях, она может быть также задействована.

Вычисление семантического сходства необходимо для установления смысловых ассоциативных отношений, позволяющих более точно классифицировать естественно-языковые тексты, а также выделять в этих текстах конкретные факты. Выделение фактов позволяет исследовать тексты на фейк-ранг (достоверность). Наиболее известными средствами вычисления сходства считаются WordNet, предлагаемый как тезаурус естественного языка с объемным хранилищем английской лексики, и Word2vec как инструментарий векторизации текстов на основе нейронных сетей. Третьим средством для вычисления сходства можно считать рассмотренную ранее BERT-модель.

Комбинированное использование перечисленных программных средств позволяет не только рассматривать семантическое сходство, но и более точно выделять родовидовые отношения, синонимию или антонимию понятий. Появляется возможность идентифицировать понятийные словарные категории и понятия, получаемые из устойчивых словосочетаний. Использование онтологических оснований приближает алгоритмы вычисления семантического сходства к семантике как средству выражения *смысловой интерпретации* текста, а не только как средству вычисления глубинно-синтаксического сходства в современных языковых ML-моделях.

В качестве примера можно привести две словарные конструкции, имеющие разную семантику при глубинно-синтаксической близости:

— "стрелять в солдат" = "война, оборона";

— "стрелять в горожан" = "террор".

Появляются семантические классы "война" и "террор". Каждый из классов индексируется уникальной меткой, которая при токенизации добавит в словарную конструкцию текста спецтокен, уточняющий семантику понятия.

Тема "террористическая атака" уже была рассмотрена ранее.

Пример кода на языке Python для инструментария word2vec выглядит следующим образом:

```
wrd2v = models.KeyedVectors.load_word2vec_
format('GoogleNews-vectors-negative300.bin',
binary = True)
word1 = 'attack'
word2 = 'shooting'
similarity = wrd2v.similarity(word1, word2)
print(f'Семантическое сходство между "{word1}" и
"{word2}" равно {similarity}')
word1 = 'attack'
word2 = 'bombing'
similarity = wrd2v.similarity(word1, word2)
print(f'Семантическое сходство между "{word1}" и
"{word2}" равно {similarity}')
```

```
< Семантическое сходство между "attack" и
  "shooting" равно 0.2511727511882782
< Семантическое сходство между "attack" и
  "bombing" равно 0.5777695775032043
```

Для языковой модели bert-base-uncased сходство вычисляется так:

```
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
model = BertModel.from_pretrained('bert-base-uncased')
# Функция определения сходства предложений
def get_bert_similarity(first_text, second_text):
    input_ids_f = tokenizer.encode(first_text, add_special_tokens = True, return_tensors = 'pt')
    input_ids_s = tokenizer.encode(second_text, add_special_tokens = True, return_tensors = 'pt')
    with torch.no_grad():
        output_f = model(input_ids_f) output_s = model(input_ids_s)
        emb_f = output_f.last_hidden_state.mean(dim = 1).squeeze().numpy()
        emb_s = output_s.last_hidden_state.mean(dim = 1).squeeze().numpy()
    similarity_sentences = 1 — cosine(emb_f, emb_s)
    print(similarity_sentences)
# Вычисление семантического сходства предложений из cnn, RT, NYTimes, TASS
sentence1 = "March 23, 2024 Shooting at Moscow concert venue leaves over 130 dead."   # cnn

sentence2 = "March 23 2024 terrorists strike at a packed concert hall in the Russian capital,
                leaving at least 60 dead."                    # RT
sentence3 = "On the morning of 11.09.2001, two Boston planes destroyed
                the World Trade Center in New York."        # NYTimes
sentence4 = "On October 11, 2022, a new multifunctional medical center was opened in Lugansk."
                                                              # TASS
get_bert_similarity(sentence1, sentence1)
get_bert_similarity(sentence1, sentence2)
get_bert_similarity(sentence1, sentence3)
get_bert_similarity(sentence1, sentence4)
    <  1
    <  0.9349522590637207
    <  0.8471251726150513
    <  0.8411691188812256
```

Из примеров видно, что предложения об актах террора в Москве и Нью-Йорке признаются семантически схожими, как и предложение об открытии многофункционального медицинского центра в Луганске.

Это означает, что при всей мощности лексических тезаурусов моделей word2vec и bert-base-uncased их современные версии оценивают скорее схожесть глубинного синтаксиса предложений, нежели реальное семантическое сходство. Такое допущение почти не оказывает влияния на использование языковых моделей в качестве классификаторов психологической окраски текста, но не может не влиять на классификацию достоверности политических новостей ("фейк — не фейк").

Становится понятным, что словарную часть существующих языковых моделей, например в части токенизации, необходимо дополнить элементами родовидовых и причинно-следственных семантических отношений, взятых из соответствующих предметной области онтологических таксономий. Ясно также, что для всех возможных предметных областей понадобится огромное число таксономий, что по трудоемкости является труднореализуемым.

Предметная область политических новостей, суженная до описания событий, связанных с военными конфликтами и террористическими атаками, не очень велика. Это дает возможность дополнить тезаурусы с относительно малыми затратами. Фрагмент такого дополнения специфических семантических категорий в русскоязычном переводе приведен в табл. 1. Пример выполнен с минимальным применением средств автоматизации, хотя дальнейшие разработки могут формировать онтологические таксономии и тезаурусы на основе дообучения языковых нейронных сетей. Механизм образцов, распознающий в тексте словарные конструкции, дает возможность не только разметить текст, но и применить эти образцы при получении тематических подборок из текстовых репозиториев для формирования обучающих корпусов и автоматического синтеза онтологий.

Приведем дословный пример публикации NYTimes.com:

The Russian authorities said on Saturday that they had arrested the four individuals suspected of setting a suburban Moscow concert on fire and killing at least 133 people, one of the worst terrorist attacks

Таблица 1

Дополнение тезаурусов для последующего использования в pattern

№	Наименование лексемы	Тип лексемы	Слова-представители
0	attacked	Образец	Атаковать, бомбить, взорвать, высаживаться, нападать, обстрелять, совершать налет, нарушить границы, нанести удары
1	begin verbs	Синсет	Высадиться, начать, открыть, создать
2	beneficiary	Образец	Выгодоприобретатели, за этим стоят, этим управляют
3	danger	Образец	Опасная ситуация, опасность
4	destroyed	Синсет	Ликвидировать уничтожить, убить
5	domination	Образец	Дискриминация, мировое господство, фашизм
6	end verbs	Синсет	Вывести, закончить, покинули, убрать
7	first start	Образец	Начать первыми
8	force verbs	Синсет	Были вынуждены, вынуждены
9	independence	Образец	Бороться за независимость, биться за освобождение, добиваться самостоятельности
10	know verbs	Синсет	Был предупрежден, знал, оповещен
11	lied	Образец	Было неправдой, солгать, слукавить, обмануть, схитрить
12	military	Синсет	Военный, война, военнослужащий, цель, оружие
13	menace	Образец	В отместку, ответить, отомстить, ответ будет, угрожать
14	struck back	Образец	Ответить на, нанести ответный удар
15	subjugate	Образец	Лишать самостоятельности, навязывать идеологию, подчинять, подминать, нарушать
16	provoked verbs	Образец	Вызвать ответ, спровоцировать
17	victims	Образец	Появились (есть, были) жертвы
18	weapon	Синсет	Артиллерия, беспилотник, гаубица, гранатомет, миномет, самолет, танк

to jolt Russia in President Vladimir V. Putin's nearly quarter century in power.

Российские власти заявили в субботу, что они арестовали четырех человек, подозреваемых в поджоге концертного зала в пригороде Москвы и убийстве по меньшей мере 133 человек, что стало одним из самых страшных террористических актов, потрясших Россию за почти четверть века пребывания у власти президента Владимира Путина.

После исключения стоп-конструкций текст будет выглядеть так:

Saturday arrested the four individuals suspected of setting a suburban Moscow concert on fire and killing at least 133 people, one of the worst terrorist attacks.

Способы сопоставления словарных конструкций этого текста с элементами рассмотренной ранее онтологии приведены в табл. 2. Здесь рассмотрен сокращенный пример разметки текста двумя спецтокенами: [EVNT] (событие) и [MLTR] (применение военного оружия).

Таблица 2

Способы сопоставления элементов онтологии

Токен	Слово	Способ сопоставления	Результат вычисления семантического сходства
[MLTR]	killing (убито)	similarity	aggress 0.51990200334701
[MLTR]	on fire (сожжен)	similarity	aggress 0.45796913146270
[MLTR]	attack (атака)	pattern	true
[EVNT]	arrest (арестовать)	similarity	retaliation 0.4425693154335022
[MLTR]	terrorist (террористический)	pattern	true
[EVNT]	suspect (подозревать)	similarity	crime 0.600471556186676

Образцы (шаблоны, patterns) для сопоставления словарных конструкций в синтаксисе Python можно записать следующим образом:

```
patterns_military = [
   [{"OP":"*"},{"ENT_TYPE": "GPE"},
   {"OP":"*"},{"LEMMA": "fire"}, {"OP":"*"}, {"LEMMA":
   "killing"}],
   [{"LEMMA": "terrorist"}, {"OP":"*"}, {"LEMMA": "at-
   tack"}],
   [{"LEMMA": "conduct"}, {}, {"LEMMA": "airstrike"}],
   [{"LEMMA": "terror"}],
   [{"LEMMA": "attack"}],
   [{"LEMMA": "kill"}]
   ]
matcher.add("MLTR", patterns_military)
patterns_event = [
   [{"LOWER": "individuals", "OP":"*"}, {"LEMMA":
   "suspected", "OP":"{1}"}],
   [{"LEMMA": "arrest"}],
   [{"LEMMA": "suspect"}]
]
matcher.add("EVNT", patterns_event)
```

Преобразованный текст со вставленными спецтокенами выглядит так:

```
<
< > Saturday [EVNT]arrested the four [EVNT] individuals [EVNT]suspected of setting a suburban Moscow concert on [MLTR]fire and [MLTR]killing at least 133 people, one of the worst [MLTR] terrorist [MLTR]attacks.
```

Для модификации bert-base-uncased токенайзера необходимо выполнить следующий фрагмент программного кода:

```
from transformers import BertTokenizer
# Загрузка предобученного токенизатора
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
# Копирование токенизатора для последующей модернизации
mod_tokenizer = tokenizer.save_pretrained("tokenizer_dir")
# Добавление спецтокенов в словарь
mod_special_tokens_dict = {'additional_special_tokens': ['[EVNT]', '[MLTR]']}
mod_tokenizer.add_special_tokens(mod_special_tokens_dict)
# Сохранение модернизированного токенизатора
mod_tokenizer.save_pretrained("tokenizer_dir")
```

5. Дообучение ML-моделей. Сравнение с известными результатами

Мониторинг англоязычных СМИ может использовать BERT-модели, изначально предобученные для решения задач, перечисленных в начале статьи. В наших исследованиях были апробированы более десятка BERT-моделей. Они проходили дообучение на корпусе из более 5000 новостных статей с задачей двоичной классификации по классам "прозападные статьи" и "пророссийские статьи", названные короткими наименованиями "Запад" и "Восток".

Характеристики наиболее эффективных моделей из репозитория "Hugging Face", дополненных моделью IYuIdeology-bert-base-uncased-ontlg, модифицированной автором статьи из известной нейросетевой языковой модели bert-based-uncased, приведены в табл. 3. Здесь все языковые модели основаны на нейронных сетях bert-based и прошли дообучение на текстовых корпусах, собранных их авторами. Для задач, которые являются для этих моделей целевыми, эти модели принадлежат группе лидеров.

Характеристики точности (F1) вычислены после дополнительного обучения на корпусе более чем из 5000 материалов СМИ.

ML-модель IYuIdeology-bert-base-uncased-correct была дообучена автором статьи на тех же входных данных на основе базовой модели bert-base-uncased. В табл. 3 приведена и характеристика F1 для модели IYuIdeology-bert-base-uncased-ontlg, которая была получена из IYuIdeology-bert-base-uncased-correct применением рассмотренной технологии онтологической токенизации.

Заключение

Экспериментальный анализ использования онтологической токенизации показал улучшение точности F1 базовой языковой модели bert-base-uncased на 2...3 % для классификационной задачи "Восток—Запад". Повышение качества моделей отмечается также в сравнении с лучшими зарубежными аналогами, оценива-

Таблица 3

Характеристики лучших по результатам исследования BERT-моделей

Название модели	Цель классификации	Время обучения, мин	Размер модели	Точность для класса Восток	Точность для класса Запад
GPA-roberta-base-openai-detector	GPT-генерация	2	490 Мбайт	0,7101	0,7190
hamzab-roberta-fake-news-classificatione	Фейк-новость	49	421 Мбайт	0,5460	0,8152
facebook-roberta-hate-speech-dynabench-r4-target	Гневный текст	2	688 Мбайт	0,6892	0,6867
martin-ha-toxic-comment-model	Токсичность	4,7	1,6 Гбайт	0,8748	0,7548
SkolkovoInstitute/russian_ toxicity_ classifier	Токсичность	2,4	702 Мбайт	*0,7060*	*0,6715*
twitter-roberta-sentiment	Эмоциональность	3,2	629 Мбайт	0,7166	0,6745
IYuIdeology-bert-base-uncased-correct	Идеология	2,2	1,3 Гбайт	0,8344	0,8110
IYuIdeology-bert-base-uncased-ontlg	Идеология	2,5	1,5 Гбайт	0,9291	0,8522

ющими психолого-эмоциональные характеристики новостных текстов СМИ (табл. 3).

При решении более сложных задач мониторинга электронных СМИ для выявления политико-идеологических мировых тенденций эффективным инструментарием являются ансамбли языковых моделей. Они позволяют прогнозировать появление и угасание очагов напряженности на основе анализа динамики публикаций.

Отслеживание дрейфа концепций проверенных языковых моделей дает возможность проследить тенденции политической эволюции руководящих органов различных государств вплоть до прогнозирования возможных межгосударственных конфликтов.

Список литературы

1. **Анастасьев А. А., Асташкин М. С., Агафонов П. А., Каширин И. Ю.** Определение достоверности новостей с использованием Ml-моделей, основанных на знаниях // IIASU'23 — Artificial intelligence in management, control, and data processing systems. Proceedings of the II All-Russian scientific conference (Moscow, April 27—28, 2023). 2023. Vol. 2. P. 21—27.
2. **Платонов Е. Н., Руденко В. Ю.** Выявление и классификация токсичных высказываний методами машинного обучения // Моделирование и анализ данных. 2022. Т. 12, № 1. С.27—48.
3. **Badjatiya P., Gupta S., Gupta M., Varma V.** Deep learning for hate speech detection in tweets // Proceedings of the 26th International Conference on World Wide Web Companion. 2017. P. 759—760.
4. **Agrawal A., An A.** Affective representations for sarcasm detection // 41st International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR 2018. P. 1029—1032.
5. **BingLiu H., Shu L., Yu Ph.S.** BERT post-training for review reading comprehension and aspect-based sentiment analysis // arXiv preprint arXiv:1904.02232 (2019).
6. **Chiarcos C., Apostol E.-S., Kabashi B., Truică C.-O.** Modelling frequency, attestation, and corpus-398 based information with OntoLex-FrAC // In Proceedings of the 29th International Conference on 400 Computational Linguistics. 2022. P. 4018—4027.
7. **Roumeliotis K. I., Tselikas N. D.** ChatGPT and Open-AI Models: A Preli-minary Review // *Future Internet*. 2023. Vol. *15*. P. 192. https: //doi.org/10.3390 /fi15060192.
8. **Международный** репозиторий для анализа данных и оригинальных технологических решений. [Электронный ресурс]. 2024. Дата обновления: 10.04.2024. URL: https://www.kaggle.com/ (дата обращения: 16.04.2022).
9. **Международный** репозиторий языковых нейросетевых моделей. [Электронный ресурс]. 2024. Дата обновления: 12.03.2024. URL: https://huggingface.co/models. (дата обращения: 26.09.2023).
10. **Каширин И. Ю.** Применение иерархической теории чисел при построении таксономии ICF для оптимизации нейронных сетей // Вестник РГРТУ. 2022. С. 118—126.
11. **The Description** Logics Handbook. Theory, Implementation and Applications. Ed. By F. Bader, D. Calvanese, D. MacGuinness, D. Nardi, P. Patel Schneider. New York: Cambridge University Press, 2003.
12. **Kashirin I. Yu., Filatov I. Yu.** Formalized Description Of Intuitive Perception Of Spatial Situations // 2019 8th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro. 2019. P. 1—4.
13. **Duineveld A. J., Stoter R., Weiden M. R., Кепера B., Benjamins V. R.** WonderTools? A comparative study of ontological engineering tools // International Journal of Human-Computer Studies. 2000. Vol. 52, N. 6. P. 1111—1133.
14. **Каширин Д. И., Каширин И. Ю., Пылькин А. Н.** Полиморфическое представление знаний в Semantic Web. М.: Горячая линия — Телеком, 2009. 138 с.
15. **Каширин И. Ю.** Иерархические числа для проектирования таксономий искусственного интеллекта ICF // Вестник РГРТУ. 2020. № 71. С.71—82.
16. **Definition** of hierarchical numbers. [Electronic resource]. 2024. Update date: 03/04/2024. URL: https://kashirin.net/definition-of-hierarchical-numbers (access date: 04/16/2022).

I. Yu. Kashirin, Dr. Tech. Sc., Professor, e-mail: igor-kashirin@mail.ru,
Ryazan State Radio Engineering University named after V. F. Utkin, Ryazan, Russian Federation,
https://orcid.org/0000-0003-1694-7410

Tokenization of Political Texts in BERT Models Using ICF^{+} Ontologies

The design of machine learning language models, as well as their ensembles, used in complex analytics of news texts of domestic and Western electronic media is considered. An example of software implementation of a new language neural network model with problem-oriented ontological tokenization is given. The language used as tools is Python v.3.10, Anaconda v.2.1. The effectiveness of the approach in comparison with the best foreign analogues is confirmed by a series of experiments using the example of classifying news articles according to their ideological orientation into Western and English-language Russian ones.

***Keywords**: Bert models, ontological models, ICF + relation, tokenizer, retriever, political news, ensembles of ML models, forecasting, semantic similarity*

DOI: 10.17587/it.30.622-632

References

1. **Anastas'yev A. A., Astashkin M. S., Agafonov P. A., Kashirin I. Yu.** Determining the reliability of news using knowledge-based ML models, *IIASU'23 — Artificial intelligence in management, control, and data processing systems. Proceedings of the II All-Russian scientific conference* (Moscow, April 27—28, 2023), 2023, vol. 2, pp. 21—27.

2. **Platonov Ye. N., Rudenko V. Yu.** Identification and classification of toxic statements by machine learning methods, *Data modeling and analysis*, 2022, vol. 12, no. 1, pp. 27—48.

3. **Badjatiya P., Gupta S., Gupta M., Varma V.** Deep learning for hate speech detection in tweets, *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 759—760.

4. **Agrawal A., An A.** Affective representations for sarcasm detection, *41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2018, pp. 1029—1032.

5. **BingLiu H., Shu L., Yu Ph. S.** BERT post-training for review reading comprehension and aspect-based sentiment analysis, *arXiv preprint* arXiv:1904.02232 (2019).

6. **Chiarcos C., Apostol E.-S., Kabashi B., Truică C.-O.** Modelling frequency, attestation, and corpus-398 based information with OntoLex-FrAC, Proceedings of the 29 th International Conference on 400 Computational Linguistics, pp. 4018—4027.

7. **Roumeliotis K. I., Tselikas N. D.** ChatGPT and Open-AI Models: A Preli-minary Review, Future Internet, 2023, vol. 15, pp. 192, available at: https: //doi.org/10.3390 /fi15060192.

8. **An international** repository for data analysis and original technological solutions. [Electronic resource]. 2024. Date of update: 10.04.2024. URL: https://www.kaggle.com/ (date of access: 16.04.2022).

9. **International** repository of language neural network models. [Electronic resource], 2024, update date: 12.03.2024, available at: https://huggingface.co/models (date of access: 26.09.2023).

10. **Kashirin I. Yu.** Application of hierarchical number theory in the construction of ICF taxonomy for optimization of neural networks, *Vestnik RGRTU*, 2022, pp. 118—126

11. **Bader F., Calvanese D., MacGuinness D., Nardi D., Patel Schneider P.** ed. The Description Logics Handbook. Theory, Implementation and Applications, New York, Cambridge University Press, 2003.

12. **Kashirin I. Yu., Filatov I. Yu.** Formalized Description of Intuitive Perception Of Spatial Situations*, 2019 8th Mediterranean Conference on Embedded Computing (MECO),* Budva, Montenegro, 2019, pp. 1—4.

13. **Duineveld A. J., Stoter R., Weiden M. R., Kenepa B., Benjamins V. R.** WonderTools? A comparative study of ontological engineering tools, *International Journal of Human-Computer Studies*, 2000, vol. 52, no. 6, pp. 1111—1133.

14. **Kashirin D. I., Kashirin I. YU., Pyl'kin A. N.** Polimorficheskoye predstavleniye znaniy v Semantic Web, Moscow, Goryachaya liniya — Telekom, 2009, 138 p.

15. **Kashirin I. Yu.** Iyerarkhicheskiye chisla dlya proyektirovaniya taksonomiy iskusstvennogo intellekta ICF, *Vestnik RGRTU*, 2020, no. 71, pp. 71—82.

16. **Definition** of hierarchical numbers. [Electronic resource], 2024, update date: 03/04/2024, available at: https://kashirin.net/definition-of-hierarchical-numbers (access date: 04/16/2022).