УДК 004.048:519.765
DOI: 10.15827/0236-235X.133.132-144

Дата подачи статьи: 10.08.20
2021. Т. 34. № 1. С. 132–144

Семантический анализ научных текстов: опыт создания корпуса и построения языковых моделей

Е.П. Бручес [1,2], *ассистент, аспирант, bruches@bk.ru*
А.Е. Паульс [1], *студент, aleksey.pauls@mail.ru*
Т.В. Батура [1,2], *к.ф.-м.н., доцент, tatiana.v.batura@gmail.com*
В.В. Исаченко [2], *аспирант, vv.isachenko@gmail.com*
Д.Р. Щербатов [1], *студент, d.shsherbatov@g.nsu.ru*

[1] *Новосибирский государственный университет, г. Новосибирск, 630090, Россия*
[2] *Институт систем информатики им. А.П. Ершова СО РАН, г. Новосибирск, 630090, Россия*

Данная статья посвящена исследованию методов автоматического обнаружения сущностей (NER) и классификации семантических отношений (RC) в научных текстах из области информационных технологий. Научные публикации содержат ценную информацию о передовых научных достижениях, однако эффективная обработка непрерывно увеличивающихся объемов данных является трудоемкой задачей. Требуется постоянное совершенствование автоматических методов обработки такой информации. Современные методы, как правило, довольно хорошо решают обозначенные задачи с помощью глубокого машинного обучения, но, чтобы добиться хорошего качества на данных из конкретных областей знаний, необходимо дообучать полученные модели на специально подготовленных корпусах. Подобные коллекции научных текстов существуют для английского языка и активно используются научным сообществом, однако в настоящее время на русском языке такие корпусы в открытом доступе не представлены.

Статья содержит описание созданного корпуса текстов на русском языке. Корпус RuSERRC состоит из 1 600 неразмеченных документов и 80 размеченных сущностями и семантическими отношениями (рассмотрены 6 типов).

В работе также предложены несколько модификаций методов для построения моделей, работающих с русским языком. Это особенно актуально, так как большая часть существующих исследований ориентирована на работу с данными на английском и китайском языках и найти в свободном доступе качественные модели для русского языка не всегда возможно. В статью включены результаты экспериментов по сравнению словарного метода, RAKE и методов на основе нейронных сетей.

Модели и корпус являются общедоступными, могут быть полезными для проведения исследований и при создании систем извлечения информации.

Ключевые слова: *распознавание сущностей, извлечение семантических отношений, классификация отношений, нейросетевые модели языка, построение корпуса текстов.*

С распространением Интернета количество информации, в том числе на естественном языке, стремительно растет. Если говорить о различных областях, то, по данным журнала «Nature», только по биомедицинской тематике мировое научное сообщество издает ежегодно свыше миллиона статей [1]. Научные публикации содержат ценную информацию о передовых научных достижениях, однако эффективная обработка столь огромных объемов данных является трудоемкой задачей. Усложняется она тем, что тексты научных статей слабоструктурированные и извлечь из них какую-либо полезную информацию нетривиально.

Предположим, стоит задача выбора научной литературы, направленной на решение некоторой проблемы определенным образом или содержащей разносторонний анализ причинно-следственных связей при описании наблюдаемого феномена. Подобный качественный отбор научной литературы в настоящее время находится за пределами возможностей стандартных поисковых систем. В связи с этим, по мнению авторов, совершенствование методов обработки информации должно быть направлено на выявление и классификацию понятий и отношений, связывающих их, чтобы можно было автоматически идентифицировать соответствующие предметно-специфические семантические отношения в научных публикациях. К примеру, было бы полезно находить и классифицировать отношения, содержащиеся в та-

ких выражениях, как «описан новый способ решения задачи» или «результаты экспериментов, полученные предложенным методом, оказались лучше ранее известных» и т.д. Выявление семантических отношений между предметно-ориентированными понятиями позволило бы выявлять исследовательские работы, посвященные той же проблеме, или отслеживать эволюцию результатов по ней.

Одной из задач извлечения информации из текстов является *распознавание именованных сущностей* (Named Entity Recognition, NER). Для ее решения необходимо найти и классифицировать упоминания именованных сущностей (слов или групп слов) в тексте по заранее определенным категориям, таким как имена людей, организации, местоположение, медицинские коды, выражения времени, денежные значения и т.д. Эта задача часто решается вместе с задачей *обнаружения отношений* (Relation Extraction, RE), суть которой состоит в выделении в тексте пар сущностей, которые могут быть связаны друг с другом. Если заранее заданы классы отношений, то говорят о задаче *классификации отношений* (Relation Classification, RC) – сопоставлении каждой паре сущностей конкретного класса отношения или распределения вероятностей классов. Для упрощения этих задач вводится следующее ограничение – сущности должны находиться в одном предложении.

Современные методы, как правило, неплохо решают обозначенные задачи с помощью глубокого машинного обучения, которое позволяет строить языковые модели на основе огромного корпуса неразмеченных текстов, например Википедии. Чтобы добиться хорошего качества на данных из конкретных областей знаний, необходимо дообучать полученные модели на специально подготовленных корпусах. В статье описан процесс создания такой коллекции текстов по информационным технологиям, которая названа RuSERRC (Russian Scientific Entity Recognition and Relation Extraction Dataset). На этом корпусе проведена серия экспериментов по исследованию и сравнению различных методов, результаты которых также включены в статью. Сам корпус, реализация методов и модели доступны по адресу https://github.com/iis-research-team.

Анализ предметной области

Задача распознавания сущностей является необходимым этапом извлечения информации из текстов. На сегодняшний день существует достаточно много датасетов общего назначения на разных языках (CONLL-2003, MUC-6, OntoNotes 5.0 и т.д.). Для решения конкретных задач, например, в биомедицинской области, используют специальные корпусы, такие как BIOCREATIVE II, BioInfer и другие. Однако для выполнения семантического анализа также важно извлекать отношения, которые связывают извлеченные сущности, поэтому стали появляться датасеты для извлечения отношений. Следует отметить, что эта задача сложнее, поскольку она неизбежно связана с семантикой языков. Вот почему имеется относительно немного доступных коллекций данных для обучения и оценки как алгоритмов распознавания именованных сущностей, так и методов извлечения и классификации отношений. Наиболее известными датасетами на английском языке общего назначения, содержащими разметку не только сущностей, но и отношений, являются CONLL04, ACE 2005, TACRED, SemEval 2010 Task 8 [2]. Для работы с русским языком наиболее популярна коллекция FactRuEval-16, состоящая из новостных текстов [3].

В последнее время появляются новые корпусы для конкретных областей знаний. Например, корпус DREC (Dutch Real Estate Classifieds) [4] на английском и голландском языках создан на основе объявлений о недвижимости. Он состоит из 2 318 документов, размеченных 9 типами сущностей и 2 типами отношений. Датасет re3d (Relationship and Entity Extraction Evaluation Dataset, https://github.com/dstl/re3d) состоит из 7 небольших наборов данных на английском языке и предназначен для анализа защиты и безопасности. Он был аннотирован с использованием двух отдельных схем для сущностей (14 типов) и отношений (11 типов). DFKI SmartData Corpus [5] состоит из 2 598 документов на немецком языке из области мобильности и промышленности и включает тексты новостных лент, сообщения в Твиттере и отчеты о трафике от радиостанций, полиции и железнодорожных компаний. Этот корпус был аннотирован 16 типами сущностей и 15 типами отношений. Недавно был опубликован корпус RuREBus на русском языке, ориентированный на решение бизнес-задач. Он состоит из большого числа неразмеченных официальных документов (распоряжений, постановлений, программ развития и планов). Размеченная часть корпуса включает в себя 188 документов и содержит разметку по 8 типам сущностей и 11 типам отношений [6].

Вызывает интерес научного сообщества и решение задач извлечения сущностей и классификации семантических отношений в научных текстах. Существуют корпусы на английском языке [7], но подобных корпусов на русском языке в открытом доступе нет.

Имеющиеся методы для решения задач NER, RE и RC можно условно разделить на три группы. Заметим, что целью этой группировки является не классификация всех этих моделей, а выделение интересных в рамках данного исследования особенностей некоторых популярных методов.

К первой группе относятся методы, которые требуют огромного количества человеческого труда, заключающегося в ручном задании правил (например, на основе регулярных выражений, деревьев синтаксического разбора и т.д.) и создании словарей для предметной области. Для этих подходов характерны высокая интерпретируемость моделей и высокое качество за счет длительной разработки и невысокой обобщающей способности. Один из таких подходов, основанный на использовании индуктивного логического программирования, описан в работе [8].

Вторая группа включает в себя методы, основанные на генерации признаков и их последующей обработке нейронными сетями (сверточными или рекуррентными) или алгоритмами классического машинного обучения (SVD-разложения). Человек отвечает за выбор этапов предобработки текста, архитектуру и подбор параметров моделей. Эти методы активно развивались до 2018 года и до сих пор применяются для решения рассматриваемых задач, так как предоставляют разумный компромисс между качеством модели, временем обучения и человеческим трудом. Описание и сравнение методов второй группы можно найти, например, в [7]. В частности, для русского языка в работе [9] предлагается использовать информацию и граф сущностей Wikidata для обучения модели, автоматического создания корпуса и сбора словаря именованных сущностей. Для задачи извлечения и классификации семантических отношений в работе [10] рекомендуется использовать так называемые синтаксические индикаторы: изначально для каждого из отношений формируется словарь лексико-синтаксических маркеров, указывающих на то или иное отношение (например, moved into для Destination, of для Component-Whole и т.д.). На вход нейронной сети подают конкатенацию изначального предложения с индикаторной фразой (e_1 indicator e_2).

Третья группа – методы, основанные на глубоком машинном обучении, в частности, на архитектурах типа Transformer. Трансформеры состоят из кодировщика и декодировщика (одна часть генерирует векторные представления, другая преобразует их в конечные данные), используют механизм self-attention (для выделения зависимостей между токенами в пределах всего предложения) и обучаются на данных большого объема, благодаря чему способны выдавать векторные представления слов, богатые синтаксической и семантической информацией. Это освобождает человека от необходимости подбора низкоуровневых признаков и позволяет сразу переходить к обнаружению и классификации семантических отношений. Модели с такой архитектурой показывают качество, сравнимое с методами из первой группы, но требуют огромного количества данных и большего времени на обработку, чем методы второй группы.

Методы третьей группы считаются наиболее перспективными на сегодняшний день. Лучший результат для задачи распознавания сущностей дает модель BERT-MRC [11] на датасете ACE 2005 для английского языка (F1-мера 86,88 %). Лучшие результаты для задачи классификации отношений показывает модель сверточной нейронной сети (EPGNN) на основе графа сущностей [12]. EPGNN объединяет семантические характеристики предложения, созданные предварительно обученной моделью BERT, с топологическими характеристиками графа. Этот метод дает макро-F1 90,2 % на датасете SemEval 2010 Task 8 и оценку микро-F1 77,1 % на коллекции ACE 2005. Еще один метод обнаружения отношений, представленный в работе [13], заключается в тонкой донастройке модели $BERT_{LARGE}$ с помощью метода сопоставления промежутков (matching the blanks). При этом модель $BERT_{LARGE}$ предварительно обучается на текстах, размеченных связанными сущностями. Этот метод дает довольно хорошие результаты на датасете SemEval 2010 Task 8 (F1-мера 89,5 %) и превосходит ранее предложенные методы на коллекции TACRED (F1-мера 71,5 %). Что касается русского языка, то одна из моделей, показывающая лучшие метрики, описана в работе [14]. В модели успешно комбинируются слои Bi-LSTM и CRF (Conditional Random Fields), что позволяет значительно увеличить качество распознавания именованных сущностей.

Разметка корпуса RuSERRC

Собранный авторами корпус состоит из находящихся в открытом доступе аннотаций научных статей по информационным технологиям. Данные были взяты из журналов «Вестник НГУ. Серия: Информационные технологии» (https://journals.nsu.ru/jit/archive/) и «Программные продукты и системы» (http://www.swsys.ru/).

Объем корпуса – 1 600 неразмеченных документов и 80 текстов, вручную размеченных сущностями и отношениями между ними. Каждый документ был размечен двумя аннотаторами независимо, разногласия были разрешены модератором. Процент согласия аннотаторов в задаче выделения сущностей составил 51,77. Значение было вычислено как отношение пересечения выделенных терминов к объединению выделенных терминов. Полученное значение показывает высокую степень субъективности при нахождении слов и фраз, являющихся терминами, и при определении точных границ сущностей, что свидетельствует о сложности решаемой задачи. Процент согласия аннотаторов при выделении и классификации отношений между сущностями составил 70,03. Значение также было вычислено как отношение пересечения выделенных отношений к объединению выделенных отношений. Более подробное описание сущностей и отношений дано далее.

В качестве сущностей рассматривались существительные или именные группы, являющиеся терминами в данной предметной области. Авторы придерживались следующего понимания термина: термин – слово или словосочетание, являющееся названием некоторого понятия определенной области науки, техники, искусства и др. В связи с бурным развитием отрасли компьютерных технологий не вся терминология в данной области устоявшаяся: одни термины появляются, другие быстро устаревают. Основной ресурс компьютерной лексики – английский язык. Для переноса терминов в русский язык применяется заимствование путем транскрипции или транслитерации, используются аббревиатуры и другие способы. Однако особую сложность представляет процедура отграничения термина от нетермина. Зачастую довольно сложно понять без контекста, является ли сочетание слов термином. Например, составной термин *модель структурной организации единого информационного пространства* в нашем корпусе считается сущностью, так как просто слово «модель» в контексте конкретной аннотации не несет полной информации. Другие примеры составных терминов: формальная модель процесса, транзакционная модель, математическая модель упругопластических сред, задачи геонавигации, задача быстрого распознавания типа устройства. В качестве ориентира при принятии решения, что можно считать сущностью, послужил тезаурус [15], первоначально задуманный как источник систематизированной информации для поддержки научной и образовательной деятельности в сфере информационных технологий. Аббревиатуры и составные термины вида «аббревиатура–термин» (ЭЭГ, ЭЭГ-данные), составные понятия, содержащие латинские символы (*n*-грамма), названия языков программирования (Python, Java, C++) или библиотек (Pytorch, Keras, pymorphy2) также было решено считать сущностями.

Разметка сущностей выполнялась в формате BIO (каждой единице текста присваивается значение тега B-TERM, если она является начальной для сущности, I-TERM, если находится внутри термина, или O, если находится вне любой сущности). В рамках такой разметки предполагается, что именованные объекты не являются рекурсивными и не перекрываются. Всего в 80 размеченных текстах содержатся 11 157 токенов и 2 027 терминов. Средняя длина термина – 2,43 слова. Самый длинный термин состоит из 11 токенов.

Классы отношений были выбраны в результате анализа работ [2, 7, 16] на основе следующих критериев:

– отношение должно толковаться однозначно (например, не рассматривается семантическое отношение Entity-Destination, так как оно также имеет косвенное значение);

– отношение должно быть способно связывать между собой научные термины (например, актантами семантического отношения Communication-Topic (an act of communication is about a topic) выступают ненаучные термины, поэтому такое отношение не подходит).

Таким образом, были выбраны шесть семантических отношений: CAUSE, COMPARE, ISA, PARTOF, SYNONYMS, USAGE. Подробное описание каждого из них приведено в таблице 1.

Отношения между сущностями выделялись только в границах одного предложения. Всего в размеченной части корпуса было выделено 620 отношений между сущностями, из них CAUSE – 25, COMPARE – 21, ISA – 90, PARTOF – 77, SYNONYMS – 22, USAGE – 385.

Таблица 1

Типы семантических отношений

Table 1

Types of semantic relations

Отношение	Описание	Примеры
CAUSE	Причинно-следственное отношение; X вызывает Y; X дает в результате Y; объект или событие дает некоторый результат	(*взаимодействие:деформация*) (*взаимодействие высокоэнергетичных пучков:деформация*) (*научные исследования:данные*)
COMPARE	Сравнение; X сравнивается с Y; X лучше/хуже Y	(*модули:ручная обработка данных*) (*реляционные БД:объектно-ориентированные*) (*ООСУБД:СУБД*)
ISA	Таксономическое отношение; отношение наследования; родо-видовые отношения; отношение между объектом и множеством, обозначающим, что объект принадлежит этому множеству; X – это Y	(*жанр:высокоуровневые характеристики*) (*импульсный нейтронный гамма-каротаж:метод*) (*Python:язык программирования*)
PARTOF	Отношение часть–целое; меронимы (обратное – холонимы); X является частью Y; Y содержит X; Y состоит из X; объект является составляющей большего целого	(*спектры гамма-излучения:сигнал*) (*приложение:Android*) (*модуль:система*)
SYNONYMS	Отношение синонимии; синонимы – слова одной части речи с полным или частичным совпадением значения	(*ЭВМ:компьютер*) (*GPU:графический процессор*) (*столбчатые диаграммы:гистограммы*)
USAGE	Отношение использования; X используется для Y; X является методом для выполнения Y; X является информацией, используемой Y	(*система:анализ генетического разнообразия*) (*WEB-технологии:программное обеспечение*) (*метод статистической обработки:анализ текстов*)

Больше половины составили отношения использования (62 %), на втором месте таксономические отношения (15 %).

Методы извлечения сущностей и классификации отношений

Словарный метод извлечения сущностей. Идея этого подхода состоит в использовании заранее составленного словаря терминов. Словарь терминов набирался в полуавтоматическом режиме двумя способами:

– из текстов научных статей были извлечены 2-, 3- и 4-граммы слов, отсортированные по значению tf-idf, затем из них вручную были выбраны фразы, которые потенциально могут быть терминами;

– из заголовков статей Википедии, входящих в подграф категории «Наука», были вручную выбраны слова и фразы, которые потенциально могут быть терминами.

Таким образом, был составлен словарь из 17 252 терминов.

Основная сложность составления словаря терминов заключается в том, что без контекста сложно понять, является фраза термином или нет. Более того, в разных контекстах одна и та же фраза может и быть, и не быть термином, например, *модель*, *текст*, *язык* и др.

Извлечение сущностей с помощью алгоритма RAKE (Rapid Automatic Keyword Extraction). Данный алгоритм предназначен для автоматического извлечения ключевых слов [17]. Сначала применяется список стоп-слов и разделителей для выделения многословных терминов. После этого используется статистическая информация: для каждого слова из ключевых фраз-кандидатов оцениваются ча-

стота, с которой оно встречалось, и количество связей между этим словом и остальными. На основании этих двух величин вычисляется вес ключевой фразы, все фразы сортируются по весам, наиболее вероятные ключевые фразы получают максимальный вес. Алгоритм хорошо применим к динамическим корпусам документов и к абсолютно новым областям знаний, при этом не зависит от языка и его особенностей.

Использовалась реализация на языке Python (https://github.com/vgrabovets/multi_rake), которая поддерживает работу с русским языком и автоматическое выделение стоп-слов из текста. Следует отметить, что при использовании заранее подготовленного списка стоп-слов из открытых источников сети Интернет (http://datalytics.ru/all/spisok-stop-slov-yandeks-direkta/) алгоритм показывает лучшие результаты.

Было замечено, что зачастую алгоритм добавляет в ключевые фразы словосочетания, содержащие глагольные формы, например, «***посвящена*** *разработке нового программного обеспечения*» или «*решения задач геонавигации* ***используются*** *алгоритмы*». Так как в качестве сущностей рассматривались существительные или именные группы, было решено выполнить предобработку текстов и убрать глаголы и их формы перед применением RAKE. Для этого использовалась обертка на питоне для инструмента mystem (https://github.com/nlpub/pymystem3) от компании Яндекс. При помощи mystem в тексте выделялись все глагольные формы и принудительно удалялись из него.

Применение CNN и LSTM для классификации отношений и извлечения сущностей. В работе [7] приведены результаты международного соревнования SemEval2018 Task 7 по извлечению и классификации семантических отношений в научных текстах на английском языке. Участниками были опробованы различные модели, отличающиеся подходом к расширению датасета, набором признаков и способами их выделения, классификатором и т.д. Исходная задача была разбита на три подзадачи: 1.1 – классификация отношений на вручную размеченных сущностях; 1.2 – классификация отношений на данных с автоматически сгенерированной разметкой сущностей; 1.3 – извлечение и классификация отношений. Авторы данной статьи рассматривают методы и результаты для подзадачи 1.1, так как она наиболее подходит к корпусу. Приводят диаграмму, отображающую популярность используемых методов (CNN, LSTM, SVM и «не использующие нейронные сети») и средний F1-score для каждого метода. Самым популярным методом оказался CNN (10 систем), по 5 систем использовали LSTM и SVM, 4 системы были построены без использования нейронных сетей. В среднем хороший результат показали решения на основе LSTM (~0.73) и CNN (~0.675), в то время как результат других решений оказался около 0,45.

Рассмотрим подробнее модель, показавшую лучший результат на подзадаче 1.1. В модели [18] были использованы эмбеддинги токенов, частеречная разметка и расстояние между сущностями в качестве признаков, в качестве классификатора – ансамбль нейронных сетей с CNN- и LSTM-блоками. Авторы предполагают, что CNN лучше выделяет кратковременные зависимости (между близко расположенными токенами), в то время как LSTM выделяет долговременные зависимости и дополняет CNN. Важную роль сыграло расширение датасета на основе генерации новых примеров с помощью языковой модели и использования перевернутых отношений (авторы утверждают, что, несмотря на нарушенные синтаксис и семантику в предложении с обратным порядком слов, некоторые зависимости между сущностями сохраняются). Также был замечен и исправлен дисбаланс в количестве обучающих примеров для разных классов отношений.

На данный момент появились модели на основе архитектуры Transformer, имеющие более глубокую структуру, использующие больше информации о языке и, как следствие, показывающие лучший результат. Несмотря на это, решение команды победителей оказалось полезным – удалось добиться хорошего результата путем грамотного расширения обучающей выборки и применения ряда трюков во время обучения, что может быть использовано и в более сложных моделях.

Классификация отношений и извлечение сущностей на основе модели BERT (Bidirectional Encoder Representations from Transformers). Модель BERT представлена компанией Google в 2018 году. На момент публикации она показала наилучшее качество для решения ряда NLP-задач.

На вход BERT принимает порядковые номера токенов в специальном словаре (например, на основе BPE), разделенные номерами специальных тэгов (меток) – [CLS] и [SEP]. Первый означает начало последовательности и аккумулирует в результате работы сети общую информацию о последовательности, второй

разделяет предложения внутри последовательности. На выходе BERT возвращает векторные представления токенов. Далее эти представления подаются на вход так называемой «голове» (BERT-head) – модели, которая решает некоторую полезную задачу на основе выходных данных BERT-а.

Для обучения BERT с нуля используются две модели – Masked LM и NSP (Next Sentence Prediction). Для этого из текстов генерируются обучающие примеры следующего вида: 15 % токенов в последовательности заменяются на тэг [MASK], а сама последовательность состоит из двух предложений, где второе либо следует в каком-то тексте за первым, либо нет (положительный и отрицательный примеры). Соответственно, цель модели Masked LM – как можно лучше восстановить замаскированные слова, а модели NSP – точно предсказать, является ли второе предложение продолжением первого. Такие задачи выбраны специально – они относятся к разным уровням оперирования семантикой языка, что позволяет выделить больше закономерностей из корпуса текстов. Имеет смысл решать эти задачи на очень большом и охватывающем корпусе текстов (например Википедии), что может занять 4 дня при использовании 64 TPU. Это существенный недостаток трансформеров, но есть способ его преодолеть, и он заключается в fine-tuning-е (дообучении).

Заметим, что для каждого языка обучать большую модель необходимо только один раз – такая модель охватывает синтаксис и семантику большинства областей языка. Для решения конкретной задачи (например, вопросно-ответная система в банковской сфере) нужно всего лишь дообучить готовую модель на относительно небольшом корпусе специфичных для задачи текстов и использовать полученную языковую модель для решения задачи. Решение конкретной задачи (NER, Sentiment Analysis (тональность текста)) заключается в дообучении BERT с новой головой (чаще всего простым классификатором).

Предобучение BERT. Для русского языка есть предобученный на русской Википедии BERT – RuBERT. Используя его веса для инициализации, можно предобучить BERT на меньшем количестве текстов из узкой предметной области. В данном случае для предобучения BERT использованы ~1 600 научных текстов на русском языке из области информационных технологий и компьютерной лингвистики.

Предобучение BERT начинается с генерации словаря предметной области. Чтобы решить проблему out-of-vocabulary случаев (когда на вход модели подается слово, которого нет в ее словаре) и ограничить число слов (что положительно скажется на информативности эмбеддингов), для создания словаря применяются специальные методы (BPE, bigram). Их основная идея – разделить все слова из некоторого корпуса текстов на маленькие части и итеративно объединять самые частотные пары соседних частей в новую часть, пока размер словаря не достигнет некоторого значения. В RuBERT максимальный размер словаря равен 119 547 токенам, часть из них зарезервированы за специальными токенами (например [CLS], [MASK]). Если фактический размер словаря меньше максимального, то неиспользованные позиции заполняются токенами вида [UNUSED X], где X – порядковый номер.

Далее будут представлены основные идеи моделей на основе BERT для NER и RC. Для упрощения в графических представлениях архитектур опущены некоторые детали, а именно функции активации (tanh, relu) и dropout. Также в моделях для NER не используется векторное представление первого токена ([CLS]).

BertLinearNer. Самый простой вариант разметки последовательностей – использовать линейный слой поверх полученных с помощью BERT векторных представлений токенов. Архитектура такой модели представлена на рисунке 1. В качестве функции потерь используется перекрестная энтропия (CrossEntropyLoss).

BertLstmLinearNer и BertCnnLinearNer. Разумным решением выглядит усложнение классификатора путем добавления слоев, обучающихся нахождению зависимостей в последовательности. Сверточные слои больше под-

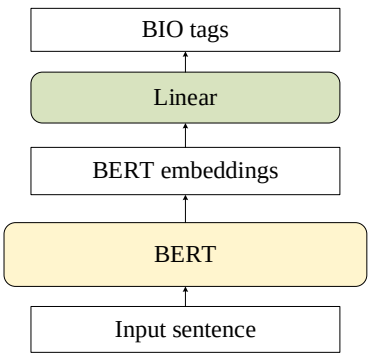

Рис. 1. BertLinearNer

Fig. 1. BertLinearNer

ходят для выделения кратковременных (для токенов на небольшом расстоянии друг от друга) зависимостей, LSTM – для выделения долговременных (для токенов на большом расстоянии друг от друга) зависимостей. В качестве структуры CNN-блоков используется структура, предложенная в проекте Keras-bert-ner (https://github.com/liushaoweihua/keras-bert-ner/). В качестве функции потерь также используется перекрестная энтропия. Полученные модели имеют архитектуру, показанную на рисунках 2 и 3.

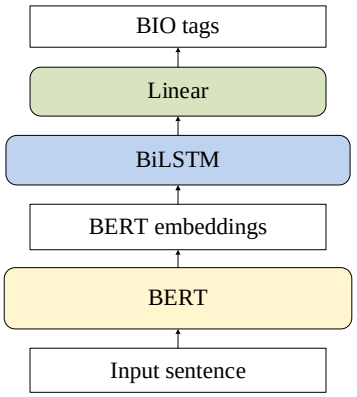

Рис. 2. BertLstmLinearNer

Fig. 2. BertLstmLinearNer

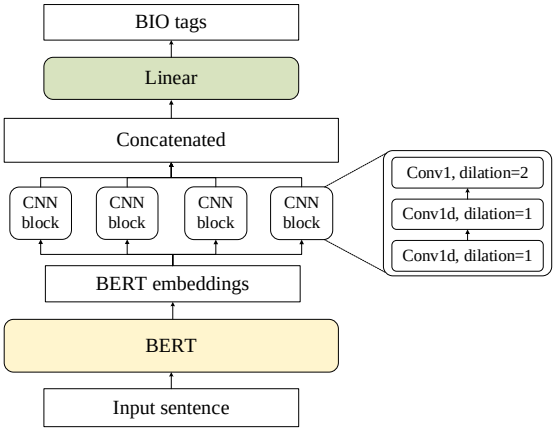

Рис. 3. BertCnnLinearNer

Fig. 3. BertCnnLinearNer

BertLstmCrfNer и BertCnnCrfNer. Следующий шаг по усложнению классификатора – добавление CRF-слоя. Вместе с CRF также используется другая функция потерь – логарифм правдоподобия. Архитектуры полученных моделей приведены на рисунках 4 и 5.

BertRC. За основу архитектуры для RC взята архитектура R-BERT [19]. Входные данные для RC отличаются: на вход, помимо последовательности токенов, также подаются битовые маски, показывающие принадлежность токенов сущностям. Это необходимо, чтобы получить из векторных представлений токенов векторные представления сущностей, например, усредняя представления токенов, входящих в сущность. Полученные векторные представления сущностей являются признаками для классификации семантического отношения. В то же время, чтобы получить векторное представление всей последовательности (всего предложения), рекомендуется брать векторное представление первого токена (специальный тег [CLS]), а не усреднять представления всех токенов. На выходе модели расположен слой Softmax, в качестве функции потерь используется MSE loss. Таким образом, была предложена архитектура, показанная на рисунке 6.

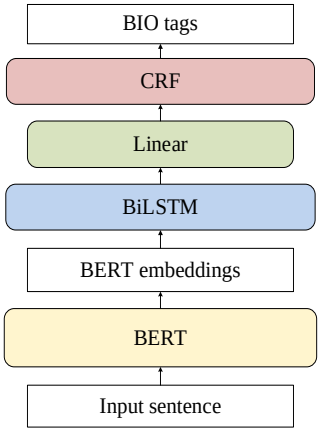

Рис. 4. BertLstmCrfNer

Fig. 4. BertLstmCrfNer

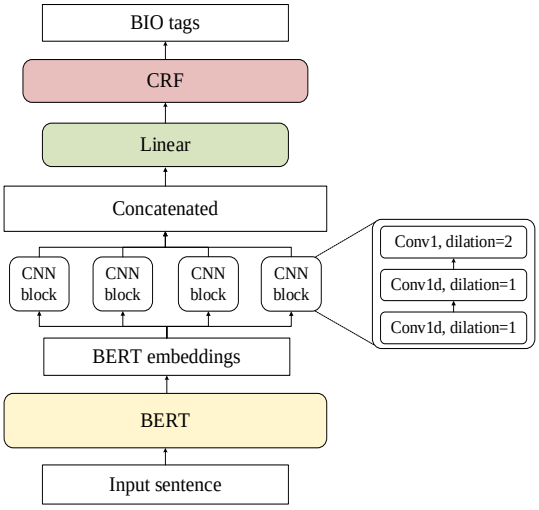

Рис. 5. BertCnnCrfNer

Fig. 5. BertCnnCrfNer

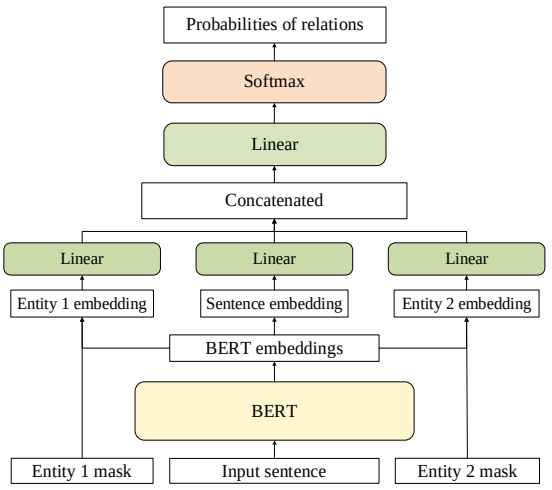

Рис. 6. BertRC

Fig. 6. BertRC

Результаты экспериментов

Для оценки качества работы алгоритмов извлечения сущностей были выбраны два вида метрик: учитывающие полное совпадение токенов и частичное. В связи с тем, что задача определения границ термина достаточно субъективна (что, например, показывает невысокая степень согласованности асессоров при разметке корпуса), метрика, учитывающая частичное совпадение токенов, видится также релевантной для данной задачи.

Точность – это доля верно извлеченных терминов относительно всех терминов, извлеченных моделью. Полнота – это доля верно найденных моделью терминов относительно всех терминов в тестовой выборке. F-мера представляет собой гармоническое среднее между точностью и полнотой.

Результаты работы вышеописанных алгоритмов извлечения сущностей на размеченных вручную данных представлены в таблице 2.

Низкие метрики у словарного метода связаны с тем, что словарь состоит только из названий терминов, в то время как в реальных текстах цепочка токенов, из которых состоит термин, может быть разорвана другими токенами, содержать синонимы, сокращения или вовсе быть неполной.

Метрики алгоритма RAKE, учитывающие полное совпадение сущностей, несколько лучше. Оптимизированная версия RAKE позволила еще немного улучшить результат. Было замечено, что алгоритм в целом выделяет больше терминов из текста, но за счет этого результат точности для метрики, учитывающей частичное совпадение токенов, ниже, чем у словарного и комбинированного подходов. При частичном совпадении сущностей оптимизация с удалением глагольных форм уменьшает мощность множества выделенных терминов, тем самым повышая точность алгоритма, но негативно влияет на полноту, так как некоторые термины все-таки содержат глагольные формы (например, причастия), и затрудняется проверка на наличие термина в размеченной выборке.

Таблица 2

Метрики нахождения сущностей

Table 2

Metrics for finding entities

Метод	Полное совпадение			Частичное совпадение		
	Точность	Полнота	F-мера	Точность	Полнота	F-мера
Словарный	0,25	0,17	0,20	**0,82**	0,34	0,48
RAKE	0,36	0,28	0,32	0,62	**0,63**	**0,63**
Оптимизированный RAKE	**0,44**	**0,35**	**0,39**	0,65	0,57	0,61

Обучение всех моделей на основе BERT проводилось в течение 50 эпох (полных проходов по набору обучающих данных). Для тестирования было выделено 10 % датасета (эти примеры не участвовали в обучении моделей). Получены следующие значения F-меры на корпусе RuSERRC для рассмотренных моделей на основе BERT: **NER**: Linear – 0,520, LSTM-Linear – **0,530**, LSTM-CRF – 0,522, CNN-Linear – 0,496, CNN-CRF – 0,503; **RC**: **0,670.**

Точность классификации отношения USAGE составляет 0,89, PARTOF – 0,59, SYNONYMS – 0,67, ISA – 0,57, COMPARE – 0,80, CAUSE – 0,50. Видно, что лучше всего модель распознает отношение USAGE, что неудивительно, так как оно самое высокочастотное в обучающей выборке. Полученные результаты могут быть улучшены при помощи расширения корпуса текстов для предобучения BERT и увеличения количества эпох для предобучения и обучения.

Ввиду того, что похожего корпуса и моделей для русского языка для решения поставленных задач авторам не удалось найти, опубликованные результаты других исследователей на других похожих датасетах могут служить лишь ориентиром и демонстрировать, какие результаты способна показать модель в принципе.

Так, для задачи RC на корпусе SemEval2018 Task7 значение F-меры составляет 0,828, на FewRel – 0,848; для задачи NER на коллекции данных Open Entity – 0,756. Безусловно, при анализе результатов нужно принимать во внимание, что значения метрик на датасете во многом зависят от его свойств – размера, полноты, качества текстов, обучающих примеров и других характеристик. Видно, что на популярных датасетах на английском языке модели, основанные на BERT и ERNIE, показывают очень хорошие результаты.

Заключение

В статье описаны эксперименты по сравнению различных методов автоматического обнаружения сущностей и классификации семантических отношений для построения моделей, работающих с русским языком. Последнее особенно актуально, так как большинство существующих исследований ориентировано на работу с данными на английском и китайском языках и найти в свободном доступе качественные модели для русского языка довольно сложно.

В будущем планируется провести ряд экспериментов на основе модели ERNIE [20], которая использует дополнительную структурированную информацию о языке. На данный момент предобученные модели ERNIE существуют только для двух языков – английского и китайского. Согласно выводам, сделанным в работе [20], есть основания полагать, что такой подход позволит улучшить результаты для русского языка.

Собранный корпус научных текстов RuSECCR, содержащий разметку сущностей и семантических отношений, является общедоступным, может быть полезен при создании систем извлечения информации и для проведения исследований.

Работа выполнена при частичной финансовой поддержке РФФИ, проект № 19-07-01134.

Литература

1. Landhuis E. Scientific literature: Information overload. Nature, 2016, vol. 535, pp. 457–458. DOI: 10.1038/nj7612-457a.
2. Hendrickx I., Kim S., Kozareva Z., Nakov P., Seaghdha D., Pade S., Pennacchiotti M., Romano L., Szpakowicz S. SemEval-2010 Task 8: Multi-Way classification of semantic relations between pairs of nominals. Proc. V Intern. Workshop on SemEval, 2010, pp. 33–38.
3. Старостин А.С., Бочаров В.В., Алексеева С.В., Бодрова А.А., Чучунков А.С., Джумаев С.С. FactRuEval 2016: Тестирование систем выделения именованных сущностей и фактов для русского языка // Матер. междунар. конф. Диалог: компьютерная лингвистика и интеллектуальные технологии. 2016. С. 688–705.
4. Bekoulis G., Deleu J., Demeester T., Develder C. Reconstructing the house from the ad: Structured prediction on real estate classifieds. Proc. XV Conf. EACL, 2017, vol. 2, pp. 274–279. DOI: 10.18653/V1/E17-2044.
5. Schiersch M., Mironova V., Schmitt M., Thomas Ph., Gabryszak A., Hennig L. A German corpus for fine-grained named entity recognition and relation extraction of traffic and industry events. Proc. LREC, 2018, pp. 4437–4444.
6. Иванин В.А., Артемова Е.Л., Батура Т.В., Иванов В.В., Саркисян В.В., Тутубалина Е.В., Смуров И.М. RuREBus-2020: Соревнование по извлечению отношений в бизнес-постановке // Матер. Междунар. конф. Диалог: компьютерная лингвистика и интеллектуальные технологии. 2020. С. 401–416.
7. Gábor K., Buscaldi D., Schumann A.-K., QasemiZadeh B., Zargayouna H., Charnois T. SemEval-2018 Task 7: Semantic relation extraction and classification in scientific papers. Proc. XII Intern. Workshop on SemEval, 2018, pp. 679–688. DOI: 10.18653/v1/S18-1111.
8. Lima R., Espinasse B., Freitas F. A logic-based relational learning approach to relation extraction: The OntoILPER system. EAAI, 2019, vol. 78, pp. 142–157.
9. Сысоев А.А., Андрианов И.А. Распознавание именованных сущностей: подход на основе вики-ресурсов // Матер. Междунар. конф. Диалог: компьютерная лингвистика и интеллектуальные технологии. 2016. С. 746–755.
10. Tao Q., Luo X., Xu R., Wang H. Enhancing relation extraction using syntactic indicators and sentential contexts. Proc. 2019 IEEE XXXI ICTAI, 2019, pp. 1574–1580. DOI: 10.1109/ICTAI.2019.00227.
11. Li X., Feng J., Meng Y., Han Q., Wu F., Li J. A unified MRC framework for named entity recognition. ArXiv, 2020, art. 11476. URL: https://arxiv.org/pdf/1910.11476.pdf (дата обращения: 12.07.2020).

12. Zhao Y., Wan H., Gao J., Lin Y. Improving relation classification by entity pair graph. JMLR: Proc. ACML, 2019, vol. 101, pp. 1156–1171.

13. Soares L.B., FitzGerald N., Ling J., Kwiatkowski T. Matching the blanks: Distributional similarity for relation learning. Proc. 57th Annual Meeting of the ACL, 2019, pp. 2895–2905. DOI: 10.18653/v1/P19-1279.

14. Anh L.T., Arkhipov M.Y., Burtsev M.S. Application of a Hybrid Bi-LSTM-CRF model to the task of russian named entity recognition. CCIS, 2018, vol. 789, pp. 91–103. DOI: 10.1007/978-3-319-71746-3_8.

15. Федотов А.М., Идрисова И.А., Самбетбаева М.А., Федотова О.А. Использование тезауруса в научно-образовательной информационной системе // Вестн. НГУ. Сер.: Информационные технологии. 2015. Т. 13. № 2. С. 86–102.

16. Aditya S., Sinha A. Uncovering relations for marketing knowledge representations. ArXiv, 2020, art. 08374. URL: https://arxiv.org/pdf/1912.08374.pdf (дата обращения: 20.07.2020).

17. Rose S., Engel D., Cramer N., Cowley W. Automatic keyword extraction from individual documents. Text Mining, 2010, pp. 1–20. DOI: 10.1002/9780470689646.CH1.

18. Rotsztejn J., Hollenstein N., Zhang C. ETH-DS3Lab at SemEval-2018. Task 7: Effectively combining recurrent and convolutional neural networks for relation classification and extraction. Proc. XII Intern. Workshop on SemEval, 2018, pp. 689–696. DOI: 10.18653/v1/S18-1112.

19. Wu S., He Y. Enriching pre-trained language model with entity information for relation classification. Proc. 28th ACM Intern. Conf. on Information and Knowledge Management, 2019, pp. 2361–2364. DOI: 10.1145/3357384.3358119.

20. Zhang Z., Han X., Liu Z., Jiang X., Sun M., Liu Q. ERNIE: Enhanced language representation with informative entities. Proc. 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1441–1451. DOI: 10.18653/v1/P19-1139.

Software & Systems
DOI: 10.15827/0236-235X.133.132-144

Received 10.08.20
2021, vol. 34, no. 1, pp. 132–144

Semantic analysis of scientific texts: Experience in creating a corpus and building language pattern

E.P. Bruches [1,2], *Assistant, Postgraduate Student, bruches@bk.ru*
A.E. Pauls [1], *Student, aleksey.pauls@mail.ru*
T.V. Batura [1,2], *Ph.D. (Physics and Mathematics), Associate Professor, tatiana.v.batura@gmail.com*
V.V. Isachenko [2], *Postgraduate Student, vv.isachenko@gmail.com*
D.R. Shsherbatov [1], *Student, d.shsherbatov@g.nsu.ru*

[1] *Novosibirsk State University, Novosibirsk, 630090, Russian Federation*
[2] *A.P. Ershov Institute of Informatics Systems (IIS), Siberian Branch of the Russian Federationn Academy of Sciences, Novosibirsk, 630090, Russian Federation*

Abstract. This paper is devoted to the development of methods for named entity recognition (NER) and relation classification (RC) in scientific texts from the information technology domain. Scientific publications provide valuable information about cutting-edge scientific advances, but efficient processing of increasing amounts of data is a time-consuming problem. Continuous improvement of automatic methods of such information processing is required. Modern deep learning methods are relatively good at solving these problems with the help of deep computer-aided learning, but in order to achieve outstanding quality on data from specific areas of knowledge, it is necessary to additional training the obtained models on the specially prepared dataset. Such collections of scientific texts are available in English and are actively used by the Russian scientific community, but at present such collections are not publicly available in Russian.

The paper contains the RuSERRC dataset description, which consists of 1600 unlabeled documents and 80 labeled with entities and semantic relations (6 relation types are considered).

Several modifications of the methods for building models for the Russian language are also proposed. This is especially important, since most of the existing research is focused on working with data in English and

Chinese, and it is not always possible to find high-quality models for the Russian language in the public domain. The paper includes the results of experiments comparing the vocabulary method, RAKE, and methods based on neural networks.

Models and datasets are publicly available, and we hope it can be useful for research purposes and the development of information extraction systems.

Keywords: named entity recognition, relation extraction, relation classification, neural network models, dataset building.

Acknowledgements. *The study was funded by RFBR according to the research project no. 19-07-01134.*

References

1. Landhuis E. Scientific literature: Information overload. *Nature*, 2016, vol. 535, pp. 457–458. DOI: 10.1038/nj7612-457a.
2. Hendrickx I., Kim S., Kozareva Z., Nakov P., Seaghdha D., Pade S., Pennacchiotti M., Romano L., Szpakowicz S. SemEval-2010 Task 8: Multi-Way classification of semantic relations between pairs of nominals. *Proc. V Intern. Workshop on SemEval*, 2010, pp. 33–38.
3. Starostin A.S., Bocharov V.V., Alexeeva S.V., Bodrova A.A., Chuchunkov A.S., Dzhumaev S.S. et al. FactRuEval 2016: Evaluation of named entity recognition and fact extraction systems for Russian. *Computational Linguistics and Intellectual Technologies: Proc. Intern. Conf. Dialogue*, 2016, pp. 688–705.
4. Bekoulis G., Deleu J., Demeester T., Develder C. Reconstructing the house from the ad: Structured prediction on real estate classifieds. *Proc. XV Conf. EACL*, 2017, vol. 2, pp. 274–279. DOI: 10.18653/V1/E17-2044.
5. Schiersch M., Mironova V., Schmitt M., Thomas Ph., Gabryszak A., Hennig L. A German corpus for fine-grained named entity recognition and relation extraction of traffic and industry events. *Proc. LREC*, 2018, pp. 4437–4444.
6. Ivanin V., Artemova E., Batura T., Ivanov V., Sarkisyan V., Tutubalina E., Smurov I. RuREBus-2020: Russian relation extraction for business. *Computational Linguistics and Intellectual Technologies: Proc. Intern. Conf. Dialogue*, 2020, pp. 401–416.
7. Gábor K., Buscaldi D., Schumann A.-K., QasemiZadeh B., Zargayouna H., Charnois T. SemEval-2018 Task 7: Semantic relation extraction and classification in scientific papers. *Proc. XII Intern. Workshop on SemEval*, 2018, pp. 679–688. DOI: 10.18653/v1/S18-1111.
8. Lima R., Espinasse B., Freitas F. A logic-based relational learning approach to relation extraction: The OntoILPER system. *EAAI*, 2019, vol. 78, pp. 142–157.
9. Sysoev A.A., Andrianov I.A. Named entity recognition in Russian: The power of wiki-based approach. *Computational Linguistics and Intellectual Technologies: Proc. Intern. Conf. Dialogue*, 2016, pp. 746–755 (in Russ.).
10. Tao Q., Luo X., Wang H. Enhancing relation extraction using syntactic indicators and sentential contexts. *Proc. 2019 IEEE XXXI ICTAI*, 2019, pp. 1574–1580. DOI: 10.1109/ICTAI.2019.00227.
11. Li X., Feng J., Meng Y., Han Q., Wu F., Li J. A unified MRC framework for named entity recognition. *ArXiv*, 2020, art. 11476. Available at: https://arxiv.org/pdf/1910.11476.pdf (accessed July 12, 2020).
12. Zhao Y., Wan H., Gao J., Lin Y. Improving relation classification by entity pair graph. *JMLR: Proc. ACML*, 2019, vol. 101, pp. 1156–1171.
13. Soares L.B., FitzGerald N., Ling J., Kwiatkowski T. Matching the blanks: Distributional similarity for relation learning. *Proc. 57th Annual Meeting of the ACL*, 2019, pp. 2895–2905. DOI: 10.18653/v1/P19-1279.
14. Anh L.T., Arkhipov M.Y., Burtsev M.S. Application of a Hybrid Bi-LSTM-CRF model to the task of russian named entity recognition. *CCIS*, 2018, vol. 789, pp. 91–103. DOI: 10.1007/978-3-319-71746-3_8.
15. Fedotov A.M., Idrisova I.A., Sambetbaeva M.A., Fedotova O.A. Using the thesaurus in the scientific and educational information system. *Vestn. NSU. Ser.: Information Technologies*, 2015, vol. 13, no. 2, pp. 86–102 (in Russ.).
16. Aditya S., Sinha A. Uncovering relations for marketing knowledge representations. *ArXiv*, 2020, art. 08374. Available at: https://arxiv.org/pdf/1912.08374.pdf (accessed July 20, 2020).
17. Rose S., Engel D., Cramer N., Cowley W. Automatic keyword extraction from individual documents. *Text Mining*, 2010, pp. 1–20. DOI: 10.1002/9780470689646.CH1.
18. Rotsztejn J., Hollenstein N., Zhang C. ETH-DS3Lab at SemEval-2018. Task 7: Effectively combining recurrent and convolutional neural networks for relation classification and extraction. *Proc. XII Intern. Workshop on SemEval*, 2018, pp. 689–696. DOI: 10.18653/v1/S18-1112.

19. Wu S., He Y. Enriching pre-trained language model with entity information for relation classification. *Proc. 28th ACM Intern. Conf. on Information and Knowledge Management*, 2019, pp. 2361–2364. DOI: 10.1145/3357384.3358119.

20. Zhang Z., Han X., Liu Z., Jiang X., Sun M., Liu Q. ERNIE: Enhanced language representation with informative entities. *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1441–1451. DOI: 10.18653/v1/P19-1139.

Для цитирования

Бручес Е.П., Паульс А.Е., Батура Т.В., Исаченко В.В., Щербатов Д.Р. Семантический анализ научных текстов: опыт создания корпуса и построения языковых моделей // Программные продукты и системы. 2021. Т. 34. № 1. С. 132–144. DOI: 10.15827/0236-235X.133.132-144.

For citation

Bruches E.P., Pauls A.E., Batura T.V., Isachenko V.V., Shsherbatov D.R. Semantic analysis of scientific texts: Experience in creating a corpus and building language pattern. *Software & Systems*, 2021, vol. 34, no. 1, pp. 132–144 (in Russ.). DOI: 10.15827/0236-235X.133.132-144.