Научная статья

УДК 004.912+004.8
DOI 10.25205/1818-7900-2022-20-3-65-76

Извлечение семантических отношений из текстов научных статей

Ольга Юрьевна Тихобаева[1], Елена Павловна Бручес[2], Татьяна Викторовна Батура[3]

[2,3]Институт систем информатики им. А. П. Ершова СО РАН
Новосибирск, Россия

[1–3]Новосибирский государственный университет
Новосибирск, Россия

[1]otikhobaeva10@gmail.com
[2]bruches@bk.ru
[3]tatiana.v.batura@gmail.com, https://orcid.org/0000-0003-4333-7888

Аннотация

В современном мире количество научных публикаций, существующих в виде электронного текста, постоянно растет. В связи с этим задачи, связанные с обработкой текстов научных статей, становятся особо актуальными. Данная работа посвящена задаче извлечения семантических отношений между сущностями из текстов научных статей на русском языке, где в качестве сущностей выступают научные термины. Извлечение отношений может быть полезно в отдельных специализированных областях, таких как поисковые и вопросно-ответные системы, а также при составлении онтологий. В ходе работы нами был создан корпус научных текстов, состоящий из 136 аннотаций научных статей на русском языке, в которых выделены 353 отношения следующих типов: USAGE, ISA, TOOL, SYNONYMS, PART_OF, CAUSE. Данный корпус использовался нами для обучения моделей. Кроме того, мы реализовали алгоритм автоматического извлечения семантических отношений и протестировали его на уже существующем корпусе научных текстов RuSERRC. Для реализации алгоритма использовалась нейросетевая модель BERT. Мы провели ряд экспериментов, связанных с использованием векторов, полученных из различных языковых моделей, а также с двумя нейросетевыми архитектурами. Разработанный инструмент и размеченный корпус выложены в открытый доступ и могут быть полезны для других исследователей.

Ключевые слова

извлечение отношений, научные термины, разметка данных, языковые модели, обработка текстов

Для цитирования

Тихобаева О. Ю., Бручес Е. П., Батура Т. В. Извлечение семантических отношений из текстов научных статей // Вестник НГУ. Серия: Информационные технологии. Т. 20, № 3. С. 65–76. DOI 10.25205/1818-7900-2022-20-3-65-76

Extracting Semantic Relations from the Texts of Scientific Articles

Olgs Y. Tikhobaeva[1], Elena P. Bruches[2], Tatyana V. Batura[3]

[2,3] A.P. Ershov Institute of Informatics Systems SB RAS
Novosibirsk, Russian Federation

[1–3]Novosibirsk State University
Novosibirsk, Russian Federation

[1]otikhobaeva10@gmail.com
[2]bruches@bk.ru
[3]tatiana.v.batura@gmail.com, https://orcid.org/0000-0003-4333-7888

Abstract

Nowadays, the number of scientific publications existing in the form of electronic text is constantly growing. As a result, the tasks related to the text processing of scientific articles become especially actual. This paper is dedicated to the task of extracting semantic relations between entities from the texts of scientific articles in Russian, where we consider scientific terms as entities. Relation extraction can be useful in some specialized areas, such as searching and question-answering systems, as well as in the compilation of ontologies. In our work, we have created a corpus of scientific texts consisting of 136 abstracts of scientific articles in Russian, in which 353 relations of the following types were highlighted: USAGE, ISA, TOOL, SYNONYMS, PART_OF, CAUSE. This corpus was used to train the machine learning models. In addition, we have implemented the automatic semantic relation extraction algorithm and tested it on the already existing corpus RuSERRC. The neural network model BERT was used to implement the algorithm. We've done a number of experiments using vectors derived from different language models, as well as two neural network architectures. The developed tool and the annotated corpus are publicly available and can be useful for other researchers.

Keywords

relation extraction, scientific terms, data annotation, language models, natural language processing

For citation

Tikhobaeva O. Yu., Bruches E. P., Batura T. V. Extracting Semantic Relations from the Texts of Scientific Articles. *Vestnik NSU. Series: Information Technologies*, 2022, vol. 20, no. 3, pp. 65–76. DOI 10.25205/1818-7900-2022-20-3-65-76

Введение

В современном мире количество научных публикаций, существующих в виде электронного текста, в частности в сети Интернет, постоянно растет. Каким образом человек может обрабатывать все эти данные? Популярная идея заключается в преобразовании неструктурированного текста в структурированный посредством разметки семантической информации. Однако большой объем и неоднородность данных делают аннотацию вручную почти невозможной. Было бы гораздо эффективнее, если бы компьютер аннотировал все данные со структурой, необходимой для решения тех или иных задач.

Очень часто в научных текстах нас интересуют именно семантические отношения между терминами. Их извлечение может быть полезно в отдельных специализированных областях, таких как поисковые и вопросно-ответные системы, а также при составлении онтологий. Именно возможности практического применения полученных структурированных данных послужили толчком к усиленному исследованию данной проблемы.

Задача извлечения семантических отношений (Relation Extraction, RE) заключается в том, что после распознавания отдельных сущностей (например, «мультимедийные технологии» и «учебный процесс») необходимо классифицировать отношения, существующие между ними (например, «USAGE (мультимедийные технологии, учебный процесс)») по контексту, в котором они находятся (например, «Показаны преимущества использования мультимедийных технологий в учебном процессе»).

На данный момент, задача остается сложной для любой предметной области, так как часто требует использования знаний вне текста (например, из баз знаний или полученных другим

путем), а также из-за отсутствия большого количества размеченных данных на русском языке для решения задачи RE.

В данной работе мы поставили перед собой две цели, во-первых, создание корпуса научных текстов с разметкой семантических отношений для русского языка, во-вторых, реализация алгоритма для извлечения семантических отношений.

Мы провели несколько экспериментов, попробовав различные подходы к проблеме извлечения семантических отношений, а затем сравнили наши результаты с результатами уже существующего алгоритма.

Обзор методов

Существуют различные подходы к решению основной задачи извлечения семантических отношений, то есть к классификации отношений между распознанными сущностями.

Самый ранний и простой из них – это подход, основанный на лексико-синтаксических шаблонах. Данный подход, подробно описанный в работе [1], заключается в выявление реальных шаблонов, которые прямо выражают те или иные семантические отношения в тексте и поиске экземпляров таких отношений при помощи этих шаблонов. Сами шаблоны могут представлять собой как просто строки, так и комбинации из частеречных тегов и лексических единиц, что характерно при работе с корпусами. Однако данный подход имеет свои недостатки, так как дает высокую точность, но низкую полноту, а также требует большого количества человеческих усилий для составления шаблонов или проверки автоматически выявленных.

Следующий подход для решения задачи извлечения семантических отношений – это метод, основанный на признаках. Данный подход можно отнести к статистическим методам извлечения отношений. Он подробно описан в работе [2]. Суть данного метода заключается в извлечении семантических и синтаксических признаков из текстов обучающих примеров, которые затем представляются классификатору в виде вектора признаков для обучения или классификации. К таким признакам авторы относят сущности, между которыми нужно установить отношение, слова между этими сущностями и их количество, типы этих сущностей, пути в дереве синтаксических связей (the syntactic parse tree) и в дереве зависимостей (the dependency tree). Основной недостаток данного подхода заключается в том, что он требует сложной синтаксической и семантической обработки текста: построения деревьев, определения типов сущностей и т. д.

Последний подход, который мы рассмотрим, это извлечение семантических отношений при помощи нейронных сетей. Исследования в этой области в основном сосредоточены на проектировании и использовании различных сетевых архитектур для захвата семантики отношений в тексте. В качестве входных данных в основном используются векторные представления слов и позиций их в тексте вместо вручную отобранных признаков. Так в работе [3] описана архитектура нейронной сети, которую авторы используют для извлечения отношений. Она включает в себя следующие три компонента: векторное представление слов, извлечение признаков и вывод. Система не нуждается в сложной синтаксической или семантической предварительной обработке данных, и на вход подаются предложения с двумя отмеченными существительными. Слова преобразуются в векторы, затем последовательно извлекаются лексические и синтаксические признаки, которые потом конкатенируются для формирования конечного вектора признаков. Наконец, для вычисления достоверности каждого отношения вектор признаков подается в softmax классификатор. Выход классификатора является вектором, размерность которого равна числу предопределенных типов отношений. Значение каждого измерения – это балл достоверности соответствующего отношения.

Описание данных

На данный момент существует корпус научных статей на русском языке RuSERRC[1] [4], который содержит 1680 текстов, 80 из которых размечены вручную, и еще 1600 не имеют разметки. В размеченных текстах выделено 4 типа информации: научные термины, семантические отношения между ними, вложенные сущности, а также связи выделенных терминов с сущностями из Wikidata. Для решения задачи RE в данном корпусе представлены 589 отношений следующих типов:

– USAGE – '*x* используется для/в *y*': 330;
– ISA – '*x* является *y*': 93;
– TOOL – '*x* позволяет создавать/изучать/и т. п. *y*': 38;
– SYNONYMS – '*x* – это то же, что *y*': 22;
– PART_OF – '*x* является частью *y*': 87;
– CAUSE – '*x* является причиной *y*': 19.

В предыдущей версии инструмента [5] применялся zero-shot подход, то есть решение задачи без предоставления материалов для обучения этому решению, а именно без обучения на русскоязычных данных. В этой работе мы попробовали улучшить качество модели за счет добавления вручную размеченных данных в обучающее множество.

Для получения таких данных мы дополнили существующий корпус RuSERRC еще 136 текстами, выделив в них 353 отношений следующих типов:

– USAGE – '*x* используется для/в *y*': 126;
– ISA – '*x* является *y*': 96;
– TOOL – '*x* позволяет создавать/изучать/и т. п.': 54;
– SYNONYMS – '*x* – это то же, что *y*': 35;
– PART_OF – '*x* является частью *y*': 23;
– CAUSE – '*x* является причиной *y*': 19.

Выделение отношений проходило следующим образом: проводился поиск выбранных для рассмотрения отношений между терминами в текстах научных статей, где ранее была проведена BIO разметка терминов. Мы рассматривали отношения между сущностями только в пределах одного предложения, и одна сущность могла участвовать в нескольких отношениях. Эти отношения были занесены в специальный файл в следующем формате: PART OF(16:18). Слово перед скобками в данной записи обозначает название типа отношений, а числа в скобках – начало первого и второго термина, между которыми установлено отношение.

Эксперименты

Извлечение отношений при помощи лексических шаблонов

Вначале мы применили подход, основанный на лексических шаблонах, указывающих на наличие того или иного отношения. Он заключается в следующем: из текстов, в которых проведена разметка терминов, извлекается контекст, находящийся между двумя терминами, проводится его лемматизация, то есть приведение к начальной форме, а затем происходит его сравнение с имеющимися лексическими шаблонами. Если находится совпадение, то между терминами устанавливается соответствующее отношение. Длина контекста не должна превышать шесть слов. Такое значение было получено экспериментально путем его изменения и сравнения качества модели. Также стоит отметить, что все лексические шаблоны были составлены вручную и включают в себя не только лексические единицы, но и знаки пунктуации. Всего для работы алгоритма используется 111 шаблонов, распределенных по типам отношений следующим образом:

– USAGE: 36;

[1] https://github.com/iis-research-team/ruserrc-dataset

– ISA: 13;
– TOOL: 29;
– SYNONYMS: 5;
– PART_OF: 5;
– CAUSE: 23.

Примеры лексических шаблонов представлены в табл. 1.

Таблица 1

Примеры лексических шаблонов

Table 1

Examples of Lexical Patterns

Отношение	**Примеры маркеров**
CAUSE	вызывает; дает в результате
ISA	является; представляет собой
PART_OF	является частью; состоит из
SYNONYMS	иначе; или
TOOL	изучает; создает
USAGE	используется для; с помощью

Однако данный подход имеет свои недостатки и сложности в работе, например, отношения могут быть не выражены в тексте лексически. Также если есть термины A-B-C, то при рассмотрении пары A-C будет выбран общий контекст, включающий и A-B, и B-C, следовательно, не ясно, какую именно пару нужно связать отношением.

Извлечение отношений, как решение стандартной задачи классификации

Следующим шагом в экспериментах стала работа с различными нейросетевыми архитектурами и языковыми моделями.

В первом варианте архитектуры мы брали предложение, спецсимволами выделяли два токена, для которых нужно определить наличие и тип отношения, затем решали задачу классификации, в которой классами выступали типы отношений и один класс для отсутствия отношения между терминами. При работе с данной архитектурой мы использовали языковую модель bert-base-multilingual-cased [6]. Это модель трансформера, предобученная на данных 104 языков с самыми большими объемами Википедии. Схема данной архитектуры представлена на рис. 1.

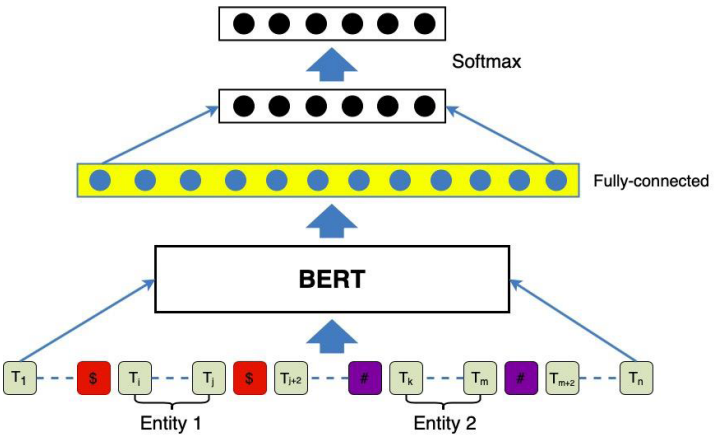

Рис. 1. Схема первой архитектуры
Fig. 1. Scheme of the first architecture

В ходе экспериментов с данной архитектурой мы пробовали обучать модель двумя разными способами. Вначале мы использовали текущую модель для извлечения отношений, обученную на англоязычных данных из корпуса SciERC [7], и дообучали ее на нашем новом русскоязычном корпусе. Затем мы попробовали обучить модель только на русскоязычных данных.

Извлечение отношений, как классификация с использованием CLS вектора

Теперь перейдем ко второму варианту нейросетевой архитектуры. Она устроена следующим образом: берется вектор специального токена CLS (считается, что он представляет собой вектор всего текста, который пришёл на вход) и вектора двух наших терминов. Затем эти три вектора конкатенируются, и полученный вектор подается в классификатор [8]. При работе с данной архитектурой мы попробовали использовать три различные языковые модели: bert-base-multilingual-cased, которая уже упоминалась ранее, rubert-base-cased от Deep Pavlov [9] и cointegrated/rubert-tiny2[2]. Модель rubert-base-cased была обучена на русскоязычной части Википедии и новостных данных. Для инициализации этой модели использовалась многоязычная версия BERT. Модель cointegrated/rubert-tiny2 – это маленькая дистиллированная версия bert-base-multilingual-cased для русского и английского языка. Она была обучена на данных Yandex Translate corpus[3], OPUS-100 [10] and Tatoeba[4]. Схема данной архитектуры представлена на рис. 2.

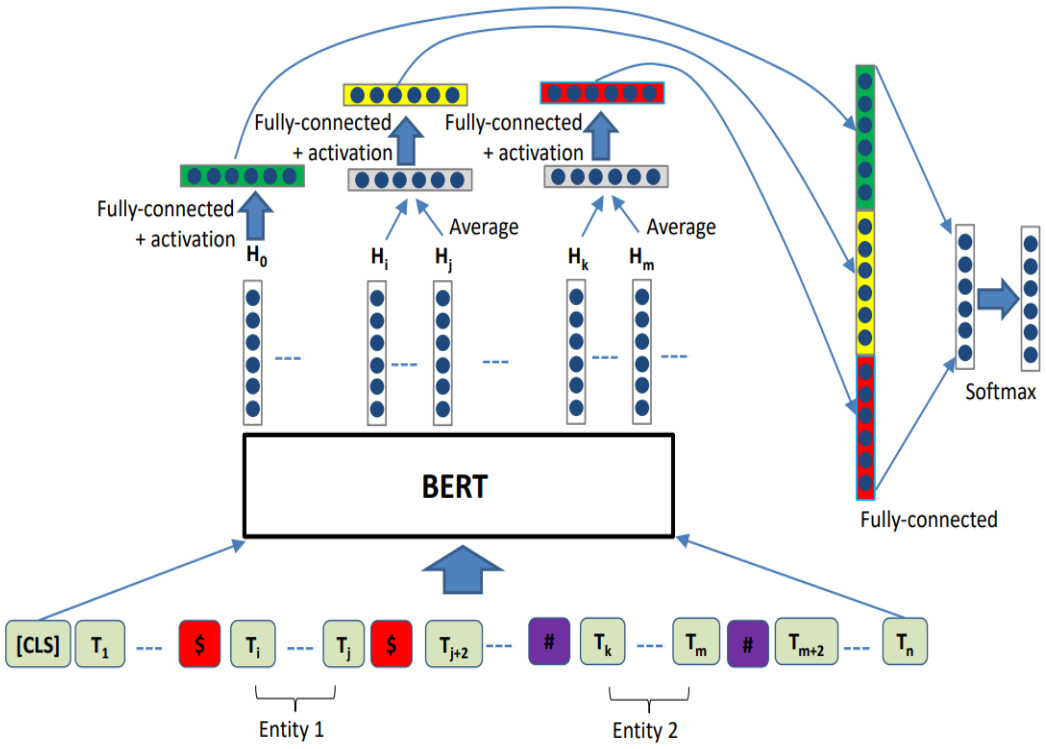

Рис. 2. Схема второй архитектуры
Fig. 2. Schema of second architecture

[2] https://huggingface.co/cointegrated/rubert-tiny2.
[3] https://translate.yandex.ru/corpus.
[4] https://tatoeba.org/ru/.

Кроме того, стоит упомянуть некоторые особенности обучения, которые заключаются в следующем: во-первых, для обучения модели мы использовали корпус русскоязычных текстов полностью, не разделяя его на обучающую и валидационную выборки, а наиболее подходящее по качеству количество эпох подбирали экспериментально. Это было сделано по причине того, что примеров некоторых отношений очень мало, и потому валидационная выборка была бы нерепрезентативной для определения качества модели; во-вторых, чтобы уменьшить дисбаланс между количеством примеров в классах, в обучающую выборку мы добавили только 50 % случайно выбранных пар терминов, исключая те, в которых расстояние между такими терминами более 10 токенов.

Результаты

Мы тестировали все варианты алгоритма на первоначальном корпусе RuSERRC, который не использовался при обучении моделей. Для оценки качества работы алгоритмов мы использовали стандартные метрики: точность, полнота, F1-мера.

По результатам работы алгоритма, основанного на лексических шаблонах, мы получили метрики, представленные в табл. 2.

Таблица 2

Метрики подхода на основе лексических шаблонов

Table 2

Metrics of Lexical Pattern's Approach

Отношение	**Точность**	**Полнота**	**F1 – мера**
CAUSE	0,07	0,05	0,06
ISA	0,18	0,19	0,19
PART_OF	0,17	0,14	0,15
SYNONYMS	0,23	0,82	0,35
TOOL	0,06	0,08	0,07
USAGE	0,21	0,39	0,27
NO-RELATION	0,96	0,92	0,94
macro-average	0,27	0,37	0,29

Эксперимент с первой архитектурой, в ходе которого мы дообучали текущую модель на новом русскоязычном корпусе, показал результаты, представленные в табл. 3. Поскольку корпуса SciERC и RuSERRC имеют разные наборы отношений, мы получали метрики только для интересующих нас отношений: HYPONYM-OF, PART-OF и USED-FOR. Данные метрики иллюстрируют комбинированный подход, при котором используются как лексические шаблоны, так и языковая модель.

Таблица 3

Метрики комбинированного подхода для дообученной модели
с англоязычными изначальными данными

Table 3

Metrics of the Combined Approach for the Pre-Trained Model
with English-Language Initial Data

Отношение	**Точность**	**Полнота**	**F1 – мера**
HYPONYM-OF	0,18	0,17	0,17
PART-OF	0,17	0,14	0,15
USED-FOR	0,16	0,39	0,23
NO-RELATION	0,96	0,9	0,93
macro-average	0,37	0,4	0,37

При попытке обучить ту же модель только на русскоязычных текстах, мы получили метрики, представленные в табл. 4. Эти метрики так же соответствуют комбинированному подходу.

Таблица 4

Метрики комбинированного подхода для модели,
обученной только на русскоязычных данных

Table 4

Metrics of the Combined Approach for the Model Trained Only in Russian-Language Data

Отношение	**Точность**	**Полнота**	**F1 – мера**
CAUSE	0,07	0,05	0,06
ISA	0,18	0,19	0,19
PART_OF	0,17	0,14	0,15
SYNONYMS	0,23	0,82	0,35
TOOL	0,06	0,08	0,07
USAGE	0,21	0,39	0,27
NO-RELATION	0,96	0,92	0,94
macro-average	0,27	0,37	0,29

При работе со второй архитектурой мы вначале получили метрики для каждой из языковых моделей. Лучшие метрики с macro-усреднением для каждой из моделей представлены в табл. 5.

Таблица 5

Метрики моделей bert-base-multilingual-cased, rubert-base-cased, rubert-tiny2

Table 5

Metrics for Bert-Base-Multilingual-Cased, Rubert-Base-Cased, Rubert-Tiny2 Models

Модель	Точность	Полнота	F1 – мера
bert-base-multilingual-cased	0,26	0,32	0,26
rubert-base-cased	**0,26**	**0,34**	**0,27**
rubert-tiny2	0,22	0,23	0,22

Затем мы получили метрики комбинированного подхода с использованием лексических шаблонов для каждой модели. Лучшие метрики с macro-усреднением для каждой из моделей представлены в табл. 6.

Таблица 6

Метрики комбинированного подхода для моделей bert-base-multilingual-cased, rubert-base-cased, rubert-tiny2

Table 6

Metrics of the Combined Approach for Bert-Base-Multilingual-Cased, Rubert-Base-Cased, Rubert-Tiny2 Models

Модель	Точность	Полнота	F1 – мера
bert-base-multilingual-cased	**0,26**	**0,41**	**0,29**
rubert-base-cased	0,29	0,35	0,28
rubert-tiny2	0,29	0,24	0,24

Лучшие метрики показал комбинированный вариант алгоритма, использующий вторую архитектуру и языковую модель bert-base-multilingual-cased. Мы сравнили результаты работы данного алгоритма с текущей версией инструмента, которая работает только с частью описанных в данной работе типов отношений. Метрики нашего алгоритма для сокращенного набора отношений, а также сравнение с текущей версией представлены в табл. 7.

Таблица 7

Метрики комбинированного подхода для модели bert-base-multilingual-cased для изначального набора отношений

Table 7

Metrics of the Combined Approach for Bert-Base-Multilingual-Cased Model with Initial Set of Relations

Отношение	Точность	Полнота	F1 – мера
HYPONYM-OF	0,27	0,37	0,31
PART-OF	0,15	0,13	0,14
USED-FOR	0,19	0,48	0,27
NO-RELATION	0,96	0,9	0,93
macro-average	**0,39**	**0,47**	**0,41**
macro-average (текущая версия инструмента)	**0,38**	**0,36**	**0,34**

Заключение

В ходе работы был создан корпус с разметкой семантических отношений для русскоязычных текстов научных статей, а также реализован алгоритм для автоматического извлечения семантических отношений. Код инструмента выложен в открытом доступе и может быть использован другими исследователями.[5]

Наши исследования показали, что для использования лексических маркеров при решении данной задачи недостаточно эксплицитного контекста, и отношения зачастую представлены в текстах неявно.

В экспериментах с использованием различных Transformer-архитектур было выявлено, что модель, использующая информацию о всём предложении (CLS-токен), существенно выигрывает в качестве.

Также мы сделали вывод, что добавление вручную размеченных данных в обучающий набор повышает общее качество системы извлечения отношений.

В дальнейшем мы планируем попробовать улучшить работу модели путем изменения параметров обучения, а также попробовать использовать другие языковые модели для сравнения результатов и нахождения лучшего варианта. Кроме того, мы собираемся расширить корпус вручную размеченных текстов, особенно уделив внимание малочисленным отношениям. Дополнительно мы хотим попробовать использовать модуль связывания сущностей для повышения качества извлечения отношений, так как в базе знаний представлены как сущности, так и их связи друг с другом, которые могут послужить основой для извлечения отношений.

Список литературы

1. **Auger A., Barrière C.** Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology*, 2008. Vol. 14, no. 1. Pp. 1–19. DOI: 10.1075/term.14.1.02 aug
2. **Kambhatla N.** Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *Proceedings of the ACL Interactive Poster and Demonstration Sessions,* 2004. pp. 178–181. DOI: 10.3115/1219044.1219066
3. **Zeng D., Liu K., Lai S., Zhou G., Zhao J.** Relation classification via convolutional deep neural network. *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, 2014. Pp. 2335–2344.
4. **Bruches E., Pauls A., Batura T., Isachenko V.** Entity recognition and relation extraction from scientific and technical texts in Russian. *2020 Science and Artificial Intelligence conference (S.A.I.ence)*, IEEE, 2020. Pp. 41–45. DOI: 10.1109/s.a.i.ence50533.2020.9303196
5. **Bruches E., Mezentseva A., Batura T.** A system for information extraction from scientific texts in Russian, 2021. arXiv preprint arXiv:2109.06703.
6. **Devlin J., Chang M.W., Lee K., Toutanova K.** Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, 2019. Vol. 1 (Long and Short Papers). Pp. 4171–4186. arXiv preprint arXiv:1810.04805. DOI: 10.18653/v1/N19-1423.
7. **Luan Y., He L., Ostendorf M., Hajishirzi H.** Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018. Pp. 3219–3232. DOI: 10.18653/v1/D18-1360

[5] https://github.com/iis-research-team/terminator#relation-extraction

8. **Wu S., He Y.** Enriching pre-trained language model with entity information for relation classification. *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019. p. 2361-2364. DOI: 10.1145/3357384.3358119
9. **Kuratov Y., Arkhipov M.** Adaptation of deep bidirectional multilingual transformers for Russian language. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*, Moscow, May 29—June 1, 2019. arXiv preprint arXiv:1905.07213
10. **Zhang B., Williams P., Titov I., Sennrich R.** Improving massively multilingual neural machine translation and zero-shot translation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. Pp. 1628–1639, Online. arXiv preprint arXiv:2004.11867. DOI: 10.18653/v1/2020.acl-main.148

References

1. **Auger A., Barrière C.** Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology*, 2008. vol. 14, no. 1, pp. 1–19. DOI: 10.1075/term.14.1.02aug
2. **Kambhatla N.** Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 2004. Pp. 178–181. DOI: 10.3115/1219044.1219066
3. **Zeng D., Liu K., Lai S., Zhou G., Zhao J.** Relation classification via convolutional deep neural network. *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, 2014. p. 2335-2344.
4. **Bruches E., Pauls A., Batura T., Isachenko V.** Entity recognition and relation extraction from scientific and technical texts in Russian. *2020 Science and Artificial Intelligence conference (S.A.I.ence)*, IEEE, 2020. Pp. 41–45. DOI: 10.1109/s.a.i.ence50533.2020.9303196
5. **Bruches E., Mezentseva A., Batura T.** A system for information extraction from scientific texts in Russian, 2021. arXiv preprint arXiv:2109.06703
6. **Devlin J., Chang M.W., Lee K., Toutanova K.** Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, 2019. Vol. 1 (Long and Short Papers), pp. 4171–4186. arXiv preprint arXiv:1810.04805. DOI: 10.18653/v1/N19-1423
7. **Luan Y., He L., Ostendorf M., Hajishirzi H.** Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018. Pp. 3219–3232. DOI: 10.18653/v1/D18-1360.
8. **Wu S., He Y.** Enriching pre-trained language model with entity information for relation classification. *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019. Pp. 2361–2364. DOI: 10.1145/3357384.3358119
9. **Kuratov Y., Arkhipov M.** Adaptation of deep bidirectional multilingual transformers for Russian language. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*, Moscow, May 29—June 1, 2019. arXiv preprint arXiv:1905.07213
10. **Zhang B., Williams P., Titov I., Sennrich R.** Improving massively multilingual neural machine translation and zero-shot translation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. Pp. 1628–1639, Online. arXiv preprint arXiv:2004.11867. DOI: 10.18653/v1/2020.acl-main.148

Информация об авторах

Тихобаева Ольга Юрьевна, студентка, Новосибирский государственный университет (Новосибирск, Россия)

Бручес Елена Павловна, младший научный сотрудник, Институт систем информатики им. А. П. Ершова СО РАН (Новосибирск, Россия); старший преподаватель, Новосибирский государственный университет (Новосибирск, Россия)

Батура Татьяна Викторовна, кандидат физико-математических наук, доцент, заведующий лабораторией, Институт систем информатики им. А. П. Ершова СО РАН (Новосибирск, Россия); доцент, Новосибирский государственный университет (Новосибирск, Россия)
ORCID: 0000-0003-4333-7888

Information about the Authors

Olga Yur. Tikhobaeva, Student, Novosibirsk State University (Novosibirsk, Russian Federation)

Elena P. Bruches, Junior Researcher, A.P. Ershov Institute of Informatics Systems SB RAS (Novosibirsk, Russian Federation); Senior Lecturer, Novosibirsk State University (Novosibirsk, Russian Federation)

Tatiana Viktorovna Batura, PhD in Physics and Mathematics, Associate Professor, Head of Laboratory, A. P. Ershov Institute of Informatics Systems SB RAS (Novosibirsk, Russian Federation); Associate Professor, Novosibirsk State University (Novosibirsk, Russian Federation)

Статья поступила в редакцию 19.05.2022;
одобрена после рецензирования 05.09.2022; принята к публикации 05.09.2022
The article was submitted 19.05.2022;
approved after reviewing 05.09.2022; accepted for publication 05.09.2022