УДК 004.912+004.8
DOI: 10.15827/0236-235X.140.698-706
Дата подачи статьи: 10.05.22, после доработки: 29.08.22
2022. Т. 35. № 4. С. 698–706

Извлечение аспектов из текстов научных статей

А.Э. Маршалова [1], *студент, a.marshalova@g.nsu.ru*
Е.П. Бручес [1, 2], *младший научный сотрудник, ст. преподаватель, bruches@bk.ru*
Т.В. Батура [2], *к.ф.-м.н., доцент, старший научный сотрудник, tatiana.v.batura@gmail.com*

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

Статья посвящена автоматическому извлечению аспектов из текстов русскоязычных научных статей. Актуальность проблемы обусловлена увеличением числа научных публикаций и возрастающей в связи с этим потребностью в автоматизированном извлечении из них основной информации и ее структурировании.

В рамках исследования был создан корпус, состоящий из 291 аннотации научных статей на русском языке, размеченных следующими аспектами: задача, цель, вклад, метод, инструмент, применение, преимущество, пример и вывод. Для каждого из выделяемых аспектов в статье приведены описания и примеры. В результате разметки корпуса были выделены 1 494 аспекта, 44 % из которых составил аспект «вклад».

В работе также предложен и реализован алгоритм автоматического извлечения аспектов из текста. Извлечение аспектов рассматривается как задача тегирования последовательности. Для реализации алгоритма используется нейронная сеть BERT. Проведен ряд экспериментов, связанных с использованием векторов, полученных из различных языковых моделей, а также с заморозкой весов модели. Лучший результат показала мультиязыковая модель, дообученная на данных авторов исследования, то есть обученная без заморозки весов. Для улучшения качества извлечения аспектов разработаны эвристики, перечисленные в статье, и произведено дообучение модели на новых данных, полученных в результате автоматической разметки с последующим ручным редактированием.

Разработанная система может быть полезна другим исследователям, так как позволяет облегчить выбор публикаций по определенной теме, обзор методов решения той или иной задачи и анализ ранее полученных результатов.

Ключевые слова*: обработка естественного языка, анализ текстовой информации, извлечение информации из текста, обработка данных, машинное обучение, нейронная сеть.*

С увеличением числа научных публикаций растет потребность в автоматизированном извлечении из них основной информации и ее структурировании, например, информации о задаче, рассматриваемой в статье, результате исследования, использованных в работе методах.

У такого рода информации нет общепризнанного названия. Например, в [1] это ключевые моменты, в работе [2] – категории научного дискурса. В данной работе такая информация названа *аспектами* статьи.

Возможность извлекать аспекты из научных статей может значительно облегчить выбор публикаций по определенной теме, обзор методов решения той или иной задачи и анализ результатов, полученных другими исследователями. Кроме того, аспекты могут использоваться для автореферирования статей [3].

Проблеме автоматического извлечения аспектов из научных статей посвящен ряд исследований. В частности, в [4] предложен подход для извлечения информации о методах: их использовании, преимуществах, альтернативных названиях и способах реализации, основанный на применении лексико-синтаксических шаблонов в сочетании со статистическим алгоритмом машинного обучения Conditional Random Fields (CRF).

В работе [2] для извлечения аспектов предлагается использовать бинарные байесовские классификаторы, количество которых соответствует количеству рассматриваемых аспектов: каждый классификатор определяет вероятность, с которой предложение относится к соответствующему аспекту.

Некоторые методы классификации предложений – Linear SVM, Random Forest, полиномиальный наивный байесовский анализ (MNB), сверточные нейронные сети, LSTM, BERT и SciBERT сравнивались в [5]. Лучший результат показала модель SciBERT.

В [6] извлечение аспектов рассматривается как задача распознавания сущностей. Под сущностью понимается совокупность всех ее упоминаний, например, Penn Treebank Tokenizer и сокращенный вариант PTB Tokenizer. Для поиска упоминаний применяется CRF-модель, обученная на векторных представлениях слов, полученных с помощью предобученной модели SciBERT и нейронной сети с архитектурой BiLSTM, после чего упоминания объединяются в сущности с помощью алгоритмов кластеризации.

Следует отметить, что описываемые в литературе методы извлечения аспектов предназначены для работы с текстами на английском языке. Кроме того, в настоящее время не существует русскоязычных корпусов научных текстов с разметкой аспектов.

Данная работа посвящена созданию корпуса русскоязычных научных текстов с ручной разметкой аспектов, а также реализации алгоритма автоматического извлечения аспектов из них.

В статье приведен список аспектов, выделяемых в текстах корпуса, и проанализированы результаты разметки. Кроме того, сделан обзор методов автоматического извлечения аспектов, подробно описаны подход, реализованный в данной работе, и метрики, полученные при тестировании предложенного алгоритма.

Создание корпуса

Формирование списка аспектов. Для формирования списка аспектов был проведен анализ литературы на предмет того, какие аспекты извлекаются другими исследователями (табл. 1). В результате выяснено, что чаще всего выделяются аспекты задача, цель, вклад, методы, результат, вывод и предпосылки работы, менее – объект исследования, гипотеза, связанная работа, новизна, модель, инструмент, данные, метрика, будущая работа. Такой набор оказался не вполне релевантным для данного исследования.

Во-первых, аспект «предпосылки» оказался трудно формализуемым. При этом он чаще всего выражен, как минимум, несколькими предложениями, в то время как выбранный авторами подход предполагает, что аспект не может выходить за границы одного предложения.

Во-вторых, понятие «вклад» включает бóльшую часть информации, относимой в некоторых работах к аспекту «результат», поэтому в последнем нет необходимости.

В-третьих, упоминания об использованных в работе инструментах обычно выделяются в тот же аспект, что и описания методов, но эти понятия следует разделять.

Наконец, анализ данных для данного исследования показал, что список можно дополнить такими аспектами, как применение, преимущество и пример.

Таблица 1

Аспекты, выделяемые в других работах

Table 1

Aspects identified in other papers

Работа	Выделяемый аспект							
	Предпосылка	**Задача**	**Цель**	**Вклад**	**Метод**	**Результат**	**Вывод**	**Другие**
Dayrell et al., 2012 [7]	+		+		+	+	+	+
Hanyurwimfura et al., 2012 [8]				+		+		
Houngubo and Mercer, 2012 [4]					+			
Liakata et al., 2012 [9]	+		+		+	+	+	
Ronzano and Saggion, 2015 [2]	+				+	+	+	+
QasemiZadeh and Schumann, 2016 [10]								+
Augenstein et al., 2017 [11]		+						+
He J. et al., 2020 [3]			+	+				+
Huang T.H.K. et al., 2020 [5]	+			+	+			
Jain S. et al., 2020 [6]		+			+			+

Таким образом, был сформирован следующий список аспектов.

1. Задача (Task) – направление, проблема, с которой связана статья: *Статья посвящена задаче <Task> рефакторинга UML-диаграмм классов </Task>*.

2. Цель (Goal) – часть задачи, которую намерены решить авторы: *Цель представленного исследования – <Goal> компьютерный поиск генов и их изоформ </Goal>*.

3. Вклад (Contrib): *<Contrib> Предложена методика поиска субоптимального разбиения </Contrib>*.

4. Метод (Method): *<Method> методом частиц-в-ячейках </Method>*.

5. Инструмент (Tool): *<Tool> система MTSS </Tool>*. Аспекты «метод» и «инструмент» похожи, так как и тот, и другой не создаются авторами, а лишь используются в работе. Однако к методам относятся последовательности действий, подходы, алгоритмы, в то время как инструмент – это то, что используется исследователями непосредственно для реализации алгоритмов: это могут быть языки программирования, фреймворки, устройства.

6. Применение (Use) – то, для чего можно использовать результаты исследования, без указания на положительные стороны работы: *Предложенный в статье подход может быть полезен при <Use> построении рекомендательных систем </Use>.*

7. Преимущество (Adv): *<Adv> полученный метод обладает сравнительно высокой точностью и быстродействием </Adv>*.

8. Пример (Example) – выбранный авторами объект для демонстрации практической части работы: *Приведен пример <Example> описания с помощью ПООП продукционных правил системы логического вывода </Example>*.

9. Вывод (Conc): *Доказано, что <Conc> любая точка этой окружности является точкой сопряжения пары круговых дуг </Conc>.*

При этом каждый аспект может быть вложенным, то есть входить в текст другого аспекта: *<Contrib> предложен базирующийся на <Tool> архиваторах </Tool> алгоритм прогнозирования временных рядов </Contrib>*.

Разметка. Созданный авторами корпус (https://github.com/iis-research-team/ruserrc-dataset/tree/master/ruserrc_aspects) состоит из аннотаций научных статей по информационным технологиям и разделен на две части, содержащие 79 и 212 текстов соответственно. Разметка выполнялась двумя асессорами, мера согласованности между которыми составила 84 %.

В ходе разметки были выделены 1 494 аспекта, при этом почти половину из них (44 %) составил аспект Contrib (табл. 2). Вероятно, это связано с тем, что корпус состоит из аннотаций, цель которых – описать проделанную авторами работу и дать представление о вкладе исследования в науку. Количество аспектов в одном тексте варьируется от 0 до 18, при этом в среднем текст содержит 5 аспектов, а один аспект состоит из 9 токенов, однако его длина во многом зависит от его типа (табл. 2).

Таблица 2

Количество и средняя длина выделенных аспектов

Table 2

Number and average length of the identified aspects

Аспект	Количество	Средняя длина в токенах
Task	186	6,3
Goal	36	12,6
Contrib	661	13,2
Method	231	3,5
Tool	106	3,0
Use	104	6,6
Adv	90	9,9
Example	36	6,7
Conc	44	16,2

Анализ длин аспектов приводит к выводу о том, что некоторые аспекты выражены терминами или короткими фразами (Method, Tool), а другие – целыми предложениями или частями сложных предложений (Goal, Contrib, Conc).

Разработка алгоритма автоматического извлечения аспектов

В рамках данной работы задача автоматического извлечения аспектов из текстов была сведена к определению для каждого токена его принадлежности к тому или иному классу, иными словами, к тегированию последовательности (sequence labeling).

Согласно разработанному алгоритму, текст сначала пропускается через предобученную языковую модель BERT [12], которая преобразует текстовые данные в векторные представления – наборы числовых признаков, характеризующих каждый токен. Затем линейный классификатор по векторному представлению каждого токена определяет, к какому аспекту он должен быть отнесен.

Схема алгоритма изображена на рисунке 1.

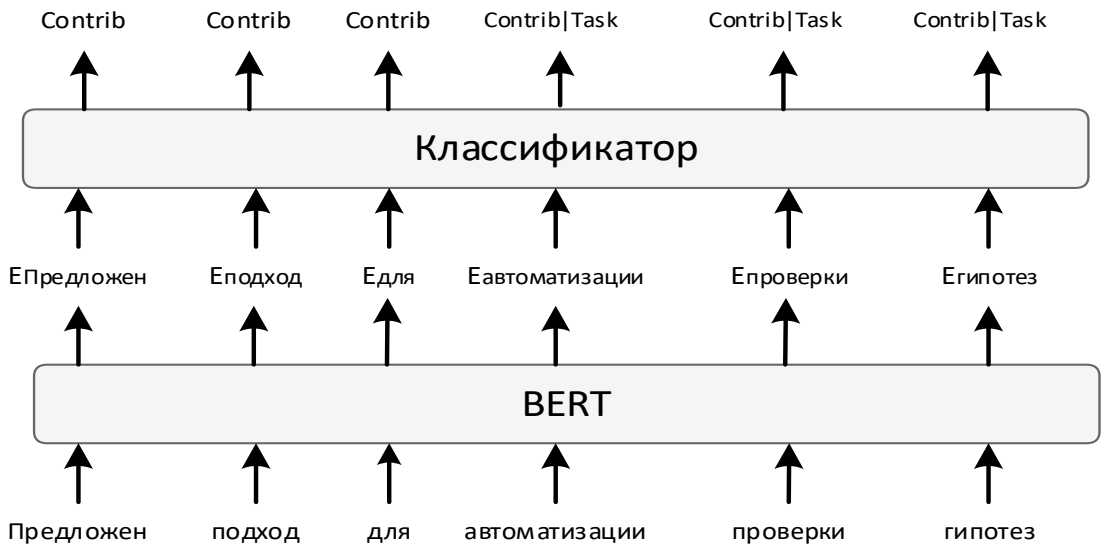

Рис. 1. Схема работы алгоритма автоматического извлечения аспектов

Fig. 1. A scheme of the automatic aspect extraction algorithm

Символом E обозначены векторные представления. На рисунке показано, что в результате классификации по их векторным представлениям, полученным с помощью BERT, все изображенные слова отнесены к аспекту Contrib, при этом слова «автоматизации», «проверки» и «гипотез» также отнесены к аспекту Task.

Алгоритм реализован на языке программирования Python с использованием сторонних библиотек tensorflow, transformers и pymorphy2. Программная реализация содержит модуль предобработки текста, модуль для извлечения аспектов и модуль постобработки извлеченных аспектов. Код программы выложен в открытый доступ (https://github.com/iis-research-team/terminator#aspect-extraction).

Преобразование данных в подходящий для задачи sequence labelling формат. Чаще всего для задачи sequence labelling используется формат BIO, то есть каждому токену присваивается тег O. Когда токен не относится ни к одному из классов, присваивается тег B-X, если токен стоит первым в последовательности, относящейся к классу X, или тег I-X, если токен входит в последовательность, относящуюся к классу X, но не является в ней первым. При такой системе тегов в два раза увеличивается количество классов, что может негативно влиять на качество предсказаний модели.

Основное преимущество этого формата в том, что он позволяет отделять друг от друга подряд идущие сущности. Однако при анализе размеченных текстов было выяснено, что аспекты одного типа никогда не следуют друг за другом подряд и разделяются, по крайней мере, запятой или союзом, поэтому нет необходимости разделения на B-теги и I-теги.

Таким образом, было предложено различать не 19, а 10 классов (девять из которых – Goal, Task, Contrib, Method, Tool, Adv, Use, Conc, Example – вышеупомянутые аспекты, а десятый – O – отсутствие аспекта).

Для учета вложенности аспектов был использован подход, предложенный в [13], согласно которому одному токену присваиваются несколько тегов (в рассматриваемом случае от одного до двух), после чего используется классификация с пересекающимися классами, где одному токену может быть присвоено более одного тега.

На рисунке 2 приведен пример текста, преобразованного в выбранный формат. Фраза «в статье» не относится ни к одному из аспектов, поэтому токенам 1 и 2 присвоен тег O, а фраза «методов аппроксимации Розенблатта–Парзена» входит одновременно в два аспекта – Contrib и Method, поэтому каждому из токенов 7–9 присвоены оба тега, разделенные вертикальной чертой.

Разработка алгоритма интерпретации предсказаний модели. Под языковой моделью понимается распределение вероятностей по по-

Id	Token	Tag
0	В	O
1	статье	O
2	приведен	Contrib
3	сравнительный	Contrib
4	анализ	Contrib
5	результатов	Contrib
6	применения	Contrib
7	методов	Contrib|Method
8	аппроксимации	Contrib|Method
9	Розенблатта–Парзена	Contrib|Method

Рис. 2. Пример размеченного текста, преобразованного в подходящий для задачи sequence labelling формат

Fig. 2. An example of an annotated text converted into a format suitable for the sequence labeling task

следовательностям токенов, то есть функция, которая принимает на вход токены и выдает на выходе вероятность присутствия каждого токена в словаре. Для подсчета вероятности каждый токен в такой модели представляется в виде вектора.

Предсказание модели – это набор вероятностей отнесения объекта к каждому из классов. Для стандартной задачи классификации (когда объект нужно отнести только к одному классу) эти вероятности вычисляются с помощью функции softmax таким образом, чтобы их сумма была равна единице, после чего выбирается наиболее вероятный класс [14]. Для классификации с пересекающимися классами используется функция sigmoid, которая определяет вероятность каждого класса вне зависимости от других классов. Затем выбираются несколько (в нашем случае до двух) самых вероятных классов при условии, что их вероятности больше определенного порогового значения, со следующими исключениями:

– если набор пуст, то есть классов, прошедших порог, не оказалось, токену присваивается класс O;

– если самый вероятный класс – O, он и присваивается токену;

– если O оказался вторым, токену присваивается только класс, оказавшийся на первом месте.

Выбор модели для обучения и получения векторных представлений. Для обучения использовалась вторая часть корпуса (212 текстов). Был проведен ряд экспериментов, связанных с использованием векторных представлений слов, полученных из различных предобученных языковых моделей: bert-base-multilingual-cased (https://huggingface.co/bert-base-multilingual-cased) [12], rubert-base-cased от DeepPavlov (https://huggingface.co/DeepPavlov/rubert-base-cased) [15], rubert-tiny2 от cointegrated (https://huggingface.co/cointegrated/rubert-tiny2), обучением модели с замороженными весами и ее дообучением на данных настоящего исследования. Лучший результат показала мультиязыковая модель bert-base-multilingual-cased, дообученная на этих данных.

Эвристики для постобработки полученных результатов. Несмотря на то, что созданная модель учитывает контекст при классификации, отнесение токенов к аспектам происходит отдельно для каждого из них, вследствие чего могут возникать ошибки, связанные с неправильным выделением границ или появлением лишних аспектов. Для устранения таких ошибок были разработаны следующие эвристики.

1. В аспект Contrib включается предшествующее ему страдательное причастие или возвратный глагол 3-го лица настоящего времени: *В статье* ***предлагается*** *<Contrib> полный минимальный список свойств, присущих интеллектуальным системам на человеческом уровне</Contrib>.* → *В статье <Contrib>* ***предлагается*** *полный минимальный список свойств, присущих интеллектуальным системам на человеческом уровне</Contrib>.*

2. Аспект не может начинаться или заканчиваться на непарный знак препинания, союз или предлог: *с использованием <Method>методов спектрального анализа </Method> <Method>* ***и метода*** *главных компонент</Method>* → *с использованием <Method> методов спектрального анализа </Method>* ***и*** *<Method>* ***метода*** *главных компонент</Method>.*

3. В аспект включается стоящая перед ним частица «не»: *разработанные модули* ***не*** *<Adv> влияют на работу других модулей </Adv>* → *разработанные модули <Adv>* ***не*** *влияют на работу других модулей </Adv>.*

4. Однословные аспекты Contrib, Conc, Goal удаляются: *на <Contrib>* ***основании*** *</Contrib> <Method> оригинального принципа компенсации первичного поля генераторной катушки </Method>* → *на* ***основании*** *<Method> оригинального принципа компенсации первичного поля генераторной катушки </Method>.*

5. Разрывы между одинаковыми аспектами удаляются: если между одинаковыми аспектами стоит другой тег и при этом сам он не является знаком препинания или союзом «и», то он заменяется на окружающие его теги, например:

а) *<Contrib> Представлено </Contrib>* ***описание*** *<Contrib> реализации ... </Contrib>* → *<Contrib> Представлено* ***описание*** *реализации ... </Contrib>*;

б) *...что обеспечило <Adv> прозрачный </Adv> <Goal>* ***доступ*** *</Goal> <Adv> к TCP-сервисам на буровой </Adv>* → *... что обеспечило <Adv> прозрачный* ***доступ*** *к TCP-сервисам на буровой </Adv>*.

Использование эвристик повысило качество извлечения аспектов (табл. 3).

Добавление новых обучающих данных. Для увеличения количества обучающих данных с помощью выбранной модели и разработанных эвристик было автоматически размечено 300 новых текстов. Впоследствии разметка отредактирована вручную.

В результате комбинированной разметки из текстов извлечено 2 810 аспектов. Модель была вновь обучена на совокупности старых (вторая часть корпуса, 212 текстов) и новых (полученных в результате полуавтоматической разметки) данных, что позволило повысить качество извлечения аспектов (табл. 3).

Результаты экспериментов

Тестирование моделей проводилось на первой части корпуса (79 текстов). Для оценки качества алгоритма использовались метрики: «точность», «полнота» и «F-1 мера» для отдельных токенов, а также «точность полного совпадения аспектов» – отношение количества полностью правильно выделенных аспектов к общему числу аспектов.

В таблице 3 представлены метрики качества отдельных токенов для исследуемых моделей.

Таблица 4 содержит метрики по каждому аспекту для лучшей модели. В работе [5] для оценки качества автоматического извлечения аспектов с помощью SciBERT приведены следующие значения F-1:

– для аспекта Finding (соответствует аспекту Contrib, выделяемому в данной работе) – 0,779;

– для аспекта Method – 0,673;

– для аспекта Purpose (соответствует аспекту Goal, выделяемому в данной работе) – 0,626.

Таблица 3

Макроусреднение по точности, полноте и F-1 для отдельных токенов для различных моделей

Table 3

Macro-average precision, recall and F-1 for individual tokens for different models

Модель	Точность	Полнота	F1
rubert-base-cased	0,174	0,200	0,178
rubert-tiny2	0,217	0,309	0,245
bert-base-multilingual-cased	0,252	0,316	0,268
bert-base-multilingual-cased + эвристики	0,272	0,307	0,276
bert-base-multilingual-cased, обученная на новых данных, + эвристики	**0,303**	**0,361**	**0,302**

Таким образом, полученные метрики оказались выше, чем в данной работе. Однако авторы использовали данные большего объема (168 286 текстов) и на английском языке.

Таблица 4

Метрики для лучшей модели

Table 4

Metrics for the best model

Тег	Точность	Полнота	F-1	Full-match accuracy
O	0,803	0,742	0,771	-
Goal	0,211	0,040	0,067	0,000
Task	0,262	0,564	0,358	0,030
Contrib	0,589	0,730	0,652	0,317
Method	0,215	0,320	0,257	0,133
Tool	0,485	0,213	0,296	0,057
Adv	0,064	0,275	0,104	0,160
Use	0,202	0,311	0,245	0,077
Conc	0,199	0,411	0,268	0,000
Example	0,000	0,000	0,000	0,000
Macro/total			0,302	0,170

Пример текста, размеченного лучшей моделью: *<Contrib> Определена модель для визуализации <Task> связей между объектами </Task> и <Task> их атрибутами в различных процессах </Task> </Contrib>. На основании модели <Contrib> разработан универсальный абстрактный компонент графического пользовательского интерфейса </Contrib> и <Contrib> приведены примеры его программной реализации </Contrib>. Также проведена*

апробация компонента для <Use> решения прикладной задачи по извлечению информации из документов </Use>.

Заключение

В ходе исследования был создан корпус русскоязычных научных текстов с ручной разметкой аспектов, разработан и реализован алгоритм автоматического извлечения аспектов из текста.

Авторы планируют продолжать эксперименты для повышения качества извлечения аспектов. Например, было замечено, что модели по-разному справляются с различными типами аспектов, поэтому ансамбль моделей видится перспективным решением для данной задачи. Кроме того, авторы планируют добавить новые эвристики, а также увеличить количество обучающих данных, так как выяснилось, что это помогает улучшить качество автоматической разметки. Наконец, результаты исследовний показали, что некоторые аспекты извлекаются плохо или не извлекаются совсем, что может быть обусловлено их неоднородностью или недостаточной представленностью в данных. В будущем необходимо пересмотреть набор аспектов и скорректировать его для того, чтобы извлекать наиболее релевантные типы информации.

Литература

1. Nasar Z., Jaffry S.W., Malik M.K. Information extraction from scientific articles: A survey. Scientometrics, 2018, vol. 117, no. 3, pp. 1931–1990. DOI: 10.1007/s11192-018-2921-5.
2. Ronzano F., Saggion H. Dr. inventor framework: Extracting structured information from scientific publications. Proc. Int. Conf. Discovery Science, 2015, pp. 209–220. DOI: 10.1007/978-3-319-24282-8_18.
3. He J., Kryściński W., McCann B., Rajani N., Xiong C. CTRLsum: Towards generic controllable text summarization. ArXiv, 2020, art. 2012.04281. URL: https://arxiv.org/abs/2012.04281 (дата обращения: 23.04.2022).
4. Houngbo H., Mercer R.E. Method mention extraction from scientific research papers. Proc. COLING, 2012, pp. 1211–1222.
5. Huang T.H.K., Huang C.Y., Ding C.K.C., Hsu Y.C., Giles C.L. CODA-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the COVID-19 open research dataset. Proc. I Workshop on NLP for COVID-19 at ACL, 2020. URL: https://arxiv.org/pdf/2005.02367v3.pdf (дата обращения: 23.04.2022).
6. Jain S., Van Zuylen M., Hajishirzi H., Beltagy I. SciREX: A challenge dataset for document-level information extraction. Proc. LVIII Annual Meeting ACL, 2020, pp. 7506–7516. DOI: 10.18653/v1/2020.acl-main.670.
7. Dayrell C., Candido Jr.A., Lima G., Machado Jr.D., Copestake A. et al. Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora. Proc. VIII Int. Conf. LREC, 2012, pp. 1604–1609.
8. Hanyurwimfura D., Liao B., Njogu H., Ndatinya E. An automated cue word based text extraction. J. of Convergence Information Technology, 2012, vol. 7, no. 10, pp. 421–429. DOI: 10.4156/JCIT.VOL7.ISSUE10.50.
9. Liakata M., Saha S., Dobnik S., Batchelor C., Rebholz-Schuhmann D. Automatic recognition of conceptualization zones in scientific articles and two life science applications. Bioinformatics, 2012, vol. 28, no. 7, pp. 991–1000. DOI: 10.1093/bioinformatics/bts071.
10. QasemiZadeh B., Schumann A.K. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. Proc. X Int. Conf. LREC, 2016, pp. 1862–1868.
11. Augenstein I., Das M., Riedel S., Vikraman L., McCallum A. SemEval 2017 task 10: ScienceIE – extracting keyphrases and relations from scientific publications. Proc. XI Int. Workshop SemEval, 2017, pp. 546–555. DOI: 10.18653/v1/S17-2091.
12. Devlin J., Chang M.W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. Proc. Conf. of the North, 2019, vol. 15, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.
13. Straková J., Straka M., Hajič J. Neural architectures for nested NER through linearization. Proc. CVII Annual Meeting of the ACL, 2019, pp. 5326–5331. DOI: 10.18653/v1/P19-1527.
14. Nwankpa C.E., Ijomah W., Gachagan A., Marshall S. Activation functions: Comparison of trends in practice and research for deep learning. Proc. II Int. Conf. on Computational Sciences and Technology, 2021, pp. 124–133.
15. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language // Computational Linguistics and Intellectual Technologies. Proc. Int. Conf. Dialogue. 2019. С. 333–339.

Software & Systems
DOI: 10.15827/0236-235X.140.698-706

Received 10.05.22, Revised 29.08.22
2022, vol. 35, no. 4, pp. 698–706

Aspect extraction from scientific paper texts

A.E. Marshalova [1], *Student, a.marshalova@g.nsu.ru*
E.P. Bruches [1, 2], *Junior Researcher, Senior Lecturer, bruches@bk.ru*
T.V. Batura [2], *Ph.D. (Physics and Mathematics), Associate Professor, Senior Researcher, tatiana.v.batura@gmail.com*

[1] *Novosibirsk State University Novosibirsk, 630090, Russian Federation*
[2] *A.P. Ershov Institute of Informatics Systems SB RAS Novosibirsk, 630090, Russian Federation*

Abstract. The paper focuses on the problem of automatic aspect extraction from the texts of Russian scientific papers. This problem is relevant due to the increase in the number of scientific publications and the growing need for automated extraction and structuring of key information from them.

The study involved the creation of a corpus consisting of 291 abstracts of Russian scientific papers annotated with the following aspects: task, goal, contribution, method, tool, use, advantage, example, and conclusion. The paper provides descriptions and examples for each aspect. As a result of the corpus annotation, 1494 aspects were identified with 44 % of them were the contribution aspect.

In addition, the paper proposes an algorithm for automatic aspect extraction. The paper considers the aspect extraction problem as a sequence-labeling problem. The BERT neural network is used to implement the algorithm. The authors have conducted a number of experiments related to the use of vectors obtained from various language models, as well as to freezing the weights of the model. A multilingual model fine-tuned on our data, that is, trained without freezing of the weights, has shown the best result. To improve the quality of aspect extraction, some heuristics, which are listed in the paper, have been developed, and the model has been further trained on the new data obtained from automatic labeling followed by manual editing.

The developed system can be useful to other researchers, as it simplifies selection of publications on a particular topic, review of methods for solving a particular problem, and analysis of results obtained in other works.

Keywords: natural language processing, text information analysis, information extraction from text, data processing, machine learning, neural networks.

References

1. Nasar Z., Jaffry S.W., Malik M.K. Information extraction from scientific articles: A survey. *Scientometrics*, 2018, vol. 117, no. 3, pp. 1931–1990. DOI: 10.1007/s11192-018-2921-5.

2. Ronzano F., Saggion H. Dr. inventor framework: Extracting structured information from scientific publications. *Proc. Int. Conf. Discovery Science*, 2015, pp. 209–220. DOI: 10.1007/978-3-319-24282-8_18.

3. He J., Kryściński W., McCann B., Rajani N., Xiong C. CTRLsum: Towards generic controllable text summarization. *ArXiv*, 2020, art. 2012.04281. Available at: https://arxiv.org/abs/2012.04281 (accessed April 23, 2022).

4. Houngbo H., Mercer R.E. Method mention extraction from scientific research papers. *Proc. COLING*, 2012, pp. 1211–1222.

5. Huang T.H.K., Huang C.Y., Ding C.K.C., Hsu Y.C., Giles C.L. CODA-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the COVID-19 open research dataset. Proc. *I Workshop on NLP for COVID-19 at ACL*, 2020. Available at: https://arxiv.org/pdf/2005.02367v3.pdf (accessed April 23, 2022).

6. Jain S., Van Zuylen M., Hajishirzi H., Beltagy I. SciREX: A challenge dataset for document-level information extraction. *Proc. LVIII Annual Meeting ACL*, 2020, pp. 7506–7516. DOI: 10.18653/v1/2020.acl-main.670.

7. Dayrell C., Candido Jr.A., Lima G., Machado Jr.D., Copestake A. et al. Rhetorical move detection in english abstracts: Multi-label sentence classifiers and their annotated corpora. *Proc. VIII Int. Conf. LREC*, 2012, pp. 1604–1609.

8. Hanyurwimfura D., Liao B., Njogu H., Ndatinya E. An automated cue word based text extraction. *J. of Convergence Information Technology*, 2012, vol. 7, no. 10, pp. 421–429. DOI: 10.4156/JCIT.VOL7.ISSUE10.50.

9. Liakata M., Saha S., Dobnik S., Batchelor C., Rebholz-Schuhmann D. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 2012, vol. 28, no. 7, pp. 991–1000. DOI: 10.1093/bioinformatics/bts071.

10. QasemiZadeh B., Schumann A.K. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. *Proc. X Int. Conf. LREC*, 2016, pp. 1862–1868.

11. Augenstein I., Das M., Riedel S., Vikraman L., McCallum A. SemEval 2017 task 10: ScienceIE – extracting keyphrases and relations from scientific publications. *Proc. XI Int. Workshop SemEval*, 2017, pp. 546–555. DOI: 10.18653/v1/S17-2091.

12. Devlin J., Chang M.W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proc. Conf. of the North*, 2019, vol. 15, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

13. Straková J., Straka M., Hajič J. Neural architectures for nested NER through linearization. *Proc. CVII Annual Meeting of the ACL*, 2019, pp. 5326–5331. DOI: 10.18653/v1/P19-1527.

14. Nwankpa C.E., Ijomah W., Gachagan A., Marshall S. Activation functions: Comparison of trends in practice and research for deep learning. *Proc. II Int. Conf. on Computational Sciences and Technology*, 2021, pp. 124–133.

15. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language. Computational Linguistics and Intellectual Technologies. *Proc. Int. Conf. Dialogue*, 2019, pp. 333–339.

Для цитирования

Маршалова А.Э., Бручес Е.П., Батура Т.В. Извлечение аспектов из текстов научных статей // Программные продукты и системы. 2022. Т. 35. № 4. С. 698–706. DOI: 10.15827/0236-235X.140.698-706.

For citation

Marshalova A.E., Bruches E.P., Batura T.V. Aspect extraction from scientific paper texts. *Software & Systems*, 2022, vol. 35, no. 4, pp. 698–706 (in Russ.). DOI: 10.15827/0236-235X.140.698-706.