Методы автоматического выявления ментальных действий в текстах научных публикаций. Часть I*

Д.А. Девяткин, Ю.М. Кузнецова, Н.В. Чудова

ИСА ФИЦ ИУ РАН, г. Москва, Россия

Аннотация. В работе показаны результаты исследования методов автоматического выделения ментальной схемы научных текстов. Рассмотрено понятие ментального действия и проанализирована проблема представленности в тексте ментальных действий и ментальных операций Изложены основания выбора схемы категоризации ментальных действий, используемой в предлагаемом подходе. Описаны разработанные методы автоматического выявления ментальных действий на основе сбора данных об использованных ментальных операциях и произведена их экспериментальная оценка.

Ключевые слова: научный текст, ментальные действия, ментальная схема научного текста, реляционно-ситуационный анализ.

DOI 10.14357/20718594180203

Введение

Развитие методов интеллектуального анализа научных текстов позволяет ставить задачи в области создания инструментов для гуманитарных и социальных исследований. Одной из таких задач является проблема построения ментальной схемы научного текста. Ее решение требует создания, прежде всего, классификатора ментальных действий и операций. В соответствии с принятым в деятельностном подходе разделении под действием понимается сознательно организованная активность, направленная на достижение поставленной цели, а под операцией – неосознаваемая активность, отвечающая условиям, в которых протекает деятельность. Для случая анализа научного текста деятельностью нужно признать познавательную активность исследователя, отражаемую им в тексте научной публикации, действиями – собственно активность по отражению различных аспектов своего исследования в научной публикации, операциями - те языковые средства, которыми автор пользуется при создании научного текста. Таким образом, с одной стороны, развитие науковедческих и психологических представлений о научной и, в целом, познавательной деятельности, может опираться на данные, полученные с помощью методов автоматического анализа текстов, а с другой, развитие средств интеллектуального анализа научных текстов связано с привлечением понятий и методов лингвистики речевых жанров и общей психологии.

1. Обзор

Одним из общих понятий для различных дисциплин и направлений, изучающих научную деятельность, является познавательное (ментальное) действие. Деятельностный подход к исследованию научного познания позволяет

*Исследование поддержано грантом РФФИ 14-29-05028 офи_м.

✉ *Девяткин Дмитрий Алексеевич E-mail: devyatkin@isa.ru*

рассматривать его как совокупность упорядоченных познавательных действий, ценностным основанием которых является стремление к объективной истине [1]. С точки зрения методологии науки последовательность познавательных действий в сочетании с нормами и правилами, применение которых приводит к достижению познавательной цели представляет собой научный метод [2]. В когнитивном освещении понятие познавательного действия наряду с понятиями субъекта и объекта познания представляют собой основные компоненты когнитивной ситуации как деятельности, направленной на познание признаков объектов окружающей действительности [3].

Лингвистика научного текста исходит из положения о том, что исследование познания возможно и дает практические результаты через изучение его репрезентации в языке (тексте) по типу реконструкции или моделирования на основе отражающих реальность языковых форм [4]. В частности, поле ментальных действий, реконструируемое с помощью лексико-семантического и компонентного анализа, включает в себя, помимо субъекта и объекта, процессы мышления и познания (выделение неизвестного), возможность и результат познания – знание, а также сохранение, передачу, контроль и коррекцию знаний и т.д. [5]. В работах Е.С. Кубряковой, В.В. Петрова, В.И. Герасимова, Р.М. Фрумкиной и др. сформулированы понятия о когнитивной грамматике как лингвистической науке о представлении знаний в языке, то есть о том, как в языковых формах отражаются процедуры приобретения человеком знаний и оперирования этими знаниями, а также результаты этих процедур – когнитивные структуры разных типов и разных уровней обобщенности. В рамках данного направления выявляются абстрактные схемы процессов познания, вербализуемых в материально данных дискурсивных структурах – фреймах со слотами «субъект познания» – «познавательное действие/отношение» – «объект познания» – «результат познания», а также таких дополнительных элементах, как «основания для квалификации», «эталонные величины», «механизмы познавательного действия» (чувственное восприятие, сравнение, сопоставление, формирование выводного знания на базе причинно-следственных отношений и т. п.), «временные и пространственные рамки ситуации», «характеристики субъекта познания» и др. Базовая структура представляет собой схему «Некто отображает (квалифицирует, интерпретирует, идентифицирует) нечто как некоторое (относящееся к какому-то классу, обладающее некоторыми свойствами, известное/неизвестное, значимое/незначимое нечто)» [4].

В рамках функционально-стилистического подхода, рассматривающего научный текст в качестве доступной для изучения формы представления знания, выработано понятие эпистемической ситуации как специфической целостности, охватывающей информацию о предмете и методе получения научного знания, а также ценностной ориентации автора. При создании научного текста для каждого аспекта эпистемической ситуации выбирается свой способ языковой номинации, что позволяет осуществлять процедуру их идентификации по текстовым маркерам лексического, пропозиционального и дискурсивного характера. Речевая реализация каждого из аспектов эпистемической ситуации получила название субтекстов, среди которых выделяются субтексты нового и старого знания, прецедентный, методологический, оценки, авторизации, адресации т.п. Важной характеристикой субтекстов является их представленность в научных произведениях в виде типизированных речевых единиц, которые в процессе развития научного стиля превращаются в стереотипные формулы, используемые в текстах различных наук и маркирующие кванты стандартного содержания [6-8]. Содержащееся в работе В.А. Салимовского [9] описание действий, образующих узловые фазы эмпирического и теоретического исследования, отличается, на наш взгляд, полнотой и системностью; для каждого действия автор приводит варианты выражающих его речевых шаблонов. Все это делает предложенную им схему анализа смыслового содержания научного текста наиболее подходящей для целей нашего исследования.

Таким образом, понятие ментальных действий (операций) на концептуальном уровне позволяет описывать строение познавательной деятельности вообще и научной в частности, а на методологическом – предоставляет широкие возможности разработки средств изучения текстовой формы существования научного знания.

Среди исследований, посвященных методам извлечения информации из текстов (information

extraction), наиболее близко к решаемой задаче стоят работы в области выявления зон аргументации (argumentative zones, AZ), в которых выделяется не менее семи категорий таких зон: формулировка цели, описание структуры глав, текущая работа (описание результатов текущих исследований), сведения общего характера (общепризнанные научные сведения), сравнение (сравнение положений статьи с другими исследованиями, описывающее слабые стороны других работ), основа (фрагменты текста, в которых автор соглашается с результатами, полученными в предшествующих работах), другие исследования (описание работ других ученых). Как можно видеть, эта схема описывает лишь наиболее значимые аспекты статьи и не в полной мере отражает структуру научной публикации. Основная задача, решаемая этой схемой – определение принадлежности декларированного в тексте знания. Однако AZ использовалась во многих работах, посвященных выделению логических зон в текстах на естественном языке, и неоднократно модифицировалась под другие предметные области и задачи [10,11].

Еще одна популярная схема – Core Scientific Concepts (CoreSC) [12] предусматривает уже 11 категорий (гипотеза, мотивация, основные положения, цель, объект исследования, метод, модель, эксперимент, наблюдения, результат, вывод) и является трехуровневой. На первом уровне описаны категории, к которым принадлежат размечаемые концепты, на втором – свойства концептов (такие как «новый», «старый», «преимущество», «недостаток»), а на третьем – уникальные идентификаторы, которые объединяют сущности, принадлежащие одному и тому же концепту. Эта схема, в отличие от предыдущей, рассматривает научную статью с точки зрения содержания исследования, однако и здесь предметом исследования является структура научной публикации, а не познавательная деятельность.

Предлагаемая нами модель основана на иных принципах. Она не просто содержит расширенный набор «риторических» или «логических» зон (не 7-11 классов, а 29 категорий), но, главное, является моделью не текста, а познавательной деятельности ученого, нашедшей свое отражение в тексте научной публикации. В качестве основы была выбрана схема ментальных действий В.А. Салимовского [9] как в наиболее полной мере отвечающей психологическим представлениям о методе операционального анализа деятельности. Проведенный нами сравнительный анализ работ, посвященных жанровой специфике научных текстов, показал, что именно схема, предложенная В.А. Салимовским, обладает наибольшей полнотой в описании активности тех или иных механизмов познавательной деятельности. Правда, действия, выделяемые В.А. Салимовскими, включают в себя только те, которые основаны на работе когнитивных механизмов, в то время как другие авторы обращают внимание и на действия, основанные на работе регулятивных механизмов психики. Поскольку никакая деятельность не может быть реализована вне взаимодействия когнитивных и регулятивных механизмов, создание полноценной модели научной (и в целом – познавательной) деятельности потребует создание объединенной схемы ментальных действий. Тем не менее, без создания средств автоматического выявления в тексте ментальных действий, опирающихся на работу таких когнитивных механизмов как ощущение, восприятие, мышление и память, в задаче моделирования познавательной деятельности не обойтись, поэтому в настоящем исследовании использовалась схема В.А. Салимовского.

Второе отличие нашего подхода является методическим. В качестве экспертов, размечающих тексты, выступали не лингвисты или программисты, а «профессиональные читатели» – специалисты в различных областях научного знания, которым, собственно говоря, и адресуются научные публикации. Таким образом, наши эксперты выступали в этом исследовании одновременно в роли испытуемых. Они читали статьи по своей специальности и, с опорой на классификатор ментальных действий, оценивали то, что сделали авторы. По механизму это было структурированное самонаблюдение (читатель наблюдал, следуя определенной схеме, за своим пониманием читаемого), по продукту – текст, размеченный в соответствии с представлением адресата научного послания о том, что делал данный адресант. Такая организация исследования позволяет нам интерпретировать полученные данные о представленности тех или иных ментальных действий и их сочетаний в научных статьях, как характеристику научной деятельности, отражаемой в публикациях, а не как характеристику текстов. Благодаря этому появляется возможность перейти от выявления

особенностей текстов, принадлежащих разным дисциплинам, научным школам, поджанрам научных публикаций, квазинаучных текстов и т.п., к описанию особенностей деятельности, порождающей эти тексты. Другими словами, интеллектуальный анализ текстов приобретает статус науковедческого и психологического инструмента.

2.Описание методов извлечения ментальных действий

Наилучшие результаты в области автоматического выявления зон аргументации в научных текстах были получены с применением методов, основанных на машинном обучении с учителем, таких как SVM (support vector machines) и условные случайные поля CRF (conditional random fields) [13, 14]. Серьезной проблемой в этой области остается вариативность текстовых средств выражения зон аргументации. Liu и др. [15] предложили в качестве их признаков использовать векторные представления (sentence embeddings), которые формируются на основе заранее обученных векторных представлений слов (word embeddings). Оценка качества предложенного подхода с использованием десятикратного перекрестного скользящего контроля, показала его преимущество по сравнению с базовыми методами. Таким образом, наиболее перспективным подходом к выявлению ментальных действий в текстах научных публикаций представляется использование методов, основанных на машинном обучении с учителем и расширение признакового пространства классифицируемых текстовых фрагментов векторными представлениями лексики, релевантной соответствующим операциям. В качестве объектов классификации в настоящей работе рассматривались предикатные слова (как единицы выражения любых операций в тексте) и наборы их аргументов. Для выявления этих объектов классификации в текстах использовался метод реляционно-ситуационного анализа [16]. В качестве признаков предикатного слова использовался идентификатор его словарной статьи и векторное представление, полученное с помощью модели Fasttext [17], предварительно обученной на текстах русскоязычной Википедии. В качестве признаков аргументов использовались идентификаторы их ролей и векторные представления, также полученные с помощью Fasttext. Векторные представления слов и словосочетаний использовались в настоящей работе с целью смягчить проблему вариативности текстовых средств выражения ментальных действий, т.е. проблему разнообразия тех ментальных операций, которые осуществляют разные авторы и распознают разные читатели. Пусть *id(p)* – номер словарной статьи предикатного слова *p*, *r(a)* – роль аргумента *a*, *w(x)* – векторное представление некоторого слова или словосочетания *x*. Тогда любое предикатное слово и его аргументы можно представить в виде:

$$Pa_i = \{\big(id(p_i), w(p_i), r(a_{i,1}), w(a_{i,1})\big), (id(p_i), w(p_i), r(a_{i,2}), w(a_{i,2}), .., (id(p_i), w(p_i), r(a_{i,n_i}), w(a_{i,n_i}))\}, \quad (1)$$

где p_i –предикатное слово, $a_{i,1}$,…,$a_{i,ni}$ – аргументы при предикатном слове p_i, $n_i \in N$ – число аргументов при предикатном слове p_i.

Таким образом, каждый текст *T* обучающего корпуса может быть представлен в виде упорядоченного множества $T = \{Pa_1, Pa_2, .., Pa_m\}$, такого что $\forall i, j < m: i < j \leftrightarrow pos(p_i, T) < pos(p_j, T)$, где $m \in N$ – число предикатных слов в тексте *T*, *pos(p,T)* – позиция предикатного слова *p* в тексте *T*.

Из-за небольшого объема и несбалансированности обучающего корпуса выявление ментальных действий в текстах производилось в два этапа. На первом этапе выявлялись классы ментальных действий, перечисленные ниже. Предварительный анализ ручной разметки ментальных действий в научных статьях [18] показал, что существует определенный порядок следования ментальных действий в текстах, который можно рассматривать как когнитивную структуру публикаций. В связи с этим задача выявления классов ментальных действий решалась с использованием методов классификации последовательностей (sequence classification), таких как условные случайные поля CRF и рекуррентные нейронные сети GRU (Gated Recurrent Unit) [19].

Архитектура нейронной сети, применяемой в методе выявления классов ментальных действий, представлена на Рис. 1, где $e_{ij} = w(p_i) + +w(a_{ij})$. Представление признаков предикатных слов и их аргументов осуществлялось в соответствии с формулой (1). Вход сети представляет собой одномерный сверточный слой, который выполняет свертки по признакам

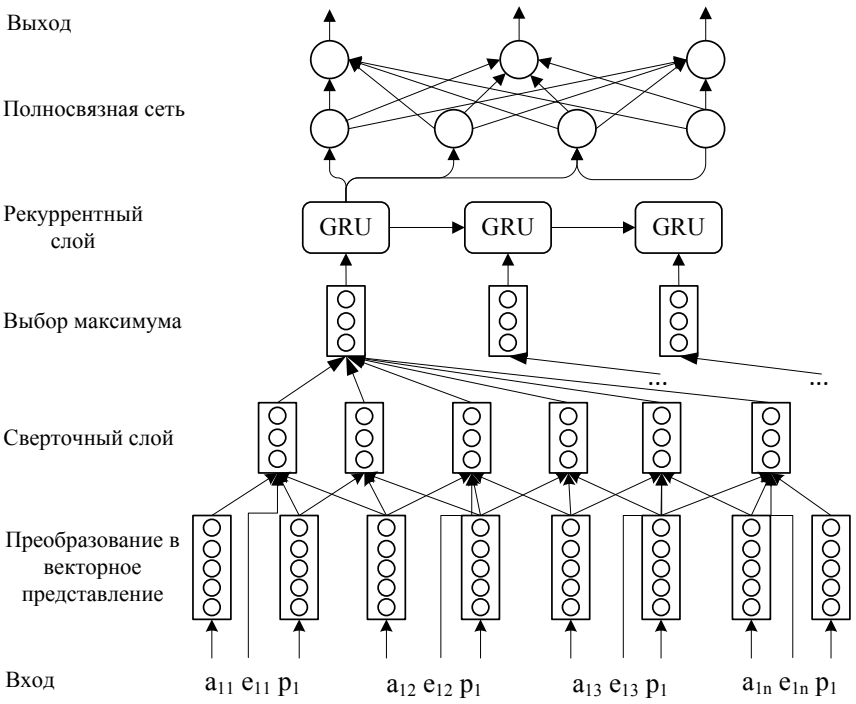

Рис. 1. Архитектура нейронной сети, применяемой в методе выявления классов ментальных действий

предикатных слов и их аргументов Pa_i. Далее следует слой для выбора максимума и рекуррентный слой GRU. Этот тип рекуррентного слоя был выбран ввиду меньшего количества параметров по сравнению с аналогами (Long-short term memory и др.), что должно снизить эффекты, связанные с переобучением. На выходе сети находится двуслойная полносвязная сеть с линейным скрытым слоем. В качестве активационной функции выходного слоя использовалась логистическая функция [20], позволяющая получать эмпирические оценки вероятностей принадлежности анализируемого фрагмента текста того или иного класса. В методе выявления классов ментальных действий строилась композиция из несколько таких нейронных сетей, причем каждая из них обучалась на своем, отобранном случайным образом подмножестве признаков. Результат работы этой композиции определялся путем голосования (soft-voting) [21]. Для решения проблемы размывания градиента (gradient vanishing), характерной для рекуррентных нейронных сетей, каждый текст разбивался на несколько пересекающихся блоков фиксированной длины, которые затем анализировались независимо друг от друга.

В методе выявления классов ментальных действий, основанном на условных случайных полях свертка признаков аргументов предикатных слов производится следующим образом:

$$Pa_i = (id(p_i), w(p_i), \mathrm{U}_{j=1}^{n_i} r(a_{i,j}), \tfrac{1}{n}\textstyle\sum_{j=1}^{n_i} w(a_{i,j})) \quad (2)$$

В ходе настоящего исследования использовалась готовая программная реализация условных случайных полей, предоставляемая библиотекой PyCRFSuite [22].

На втором этапе выявлялись собственно ментальные действия, о существовании которых в познавательной деятельности автора мы судим по выполненным им при создании научного текста ментальным действиям. Для выявления ментальных действий было предложено два метода: на основе случайного леса деревьев решений (Random Forest) [23] и на основе сверточной нейронной сети. Архитектура нейронной сети (Рис. 2) для выявления ментальных действий схожа с архитектурой сети, применяемой в методе выявления класса ментальных действий, однако в ней отсутствует рекуррентный слой.

В методе выявления ментальных действий на основе случайного леса деревьев решений использовалась готовая программная реализация из библиотеки scikit-learn [24]. Представление признаков предикатных слов и их аргументов осуществлялось в соответствии с формулой (2).

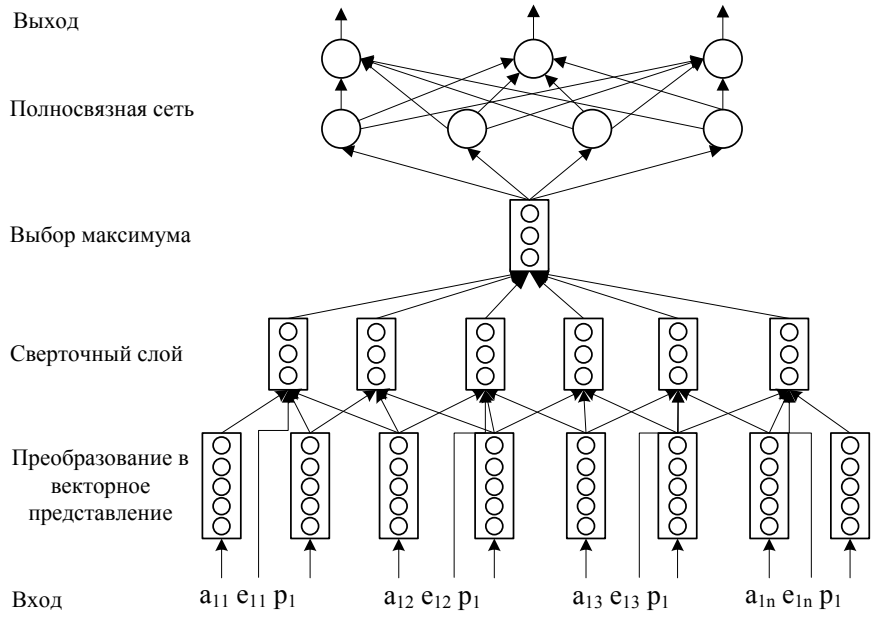

Рис. 2. Архитектура нейронной сети, применяемой для выявления ментальных действий

3. Описание размеченного корпуса

В размеченный корпус вошло 102 научные статьи на русском языке из следующих областей науки: физика, химия, психология, социология, культурология, информационные технологии, искусствоведение. Разметка корпуса производилась экспертами, обладающими учеными степенями в этих областях. Результаты разметки проверялись специалистами в области психолингвистики. Классы ментальных действий распределены в этом корпусе неравномерно (Табл. 1), что затрудняет обучение машинных методов выявления ментальных действий на основе выполненных авторами и читателями ментальных операций.

Табл. 1. Распределение классов ментальных действий в размеченном корпусе

Класс	**Количество фрагментов**
Описание нового для науки явления	1711
Сообщение об эмпирической закономерности причинно-следственного типа	2155
Классификация: представление результатов классификации данных опыта	771
Определение принципов разрабатываемой теории	851
Критический анализ наличного теоретического знания	3840
Проверка теории экспериментальным методом	6093

4. Оценка методов извлечения ментальных действий на размеченном корпусе

Оценка качества методов выявления классов ментальных действий выполнялась на размеченном корпусе с помощью статистической процедуры скользящего контроля [25]. Использовались стандартные для метрики оценки качества методов, основанных на машинном обучении с учителем, такие как точность P, полнота R и F_1 – мера [26]. Полученные оценки качества для метода, основанного на условном случайном поле (CRF) и на композиции нейронных сетей (NN_1), представлены в Табл. 2.

Из Табл. 2 видно, что метод на основе композиции нейронных сетей позволяет добиться более высоких оценок качества выявления классов ментальных действий, чем CRF. Необходимо также отметить, что получены достаточно низкие оценки качества выявления классов ментальных действий. Для повышения качества в ходе дальнейших исследований необходимо расширить и сбалансировать имеющийся размеченный корпус, а также согласовать разметку, представленную экспертами.

Табл. 2. Оценки качества выявления классов ментальных действий

	CRF			**NN_1**		
Гиперкатегория	P	R	F_1	P	R	F_1
Проверка теории экспериментальным методом	0.45±0.03	0.47±0.02	0.46±0.02	0.48±0.01	0.56±0.02	**0.52±0.01**
Критический анализ наличного теоретического знания	0.58±0.02	0.80±0.01	0.67±0.01	0.62±0.01	0.79±0.02	**0.70±0.01**
Определение принципов разрабатываемой теории	0.19±0.01	0.14±0.01	**0.16**±0.01	0.29±0.03	0.07±0.02	0.12±0.03
Описание нового для науки явления	0.25±0.01	0.12± 0.02	0.16±0.02	0.30±0.02	0.17±0.01	**0.21±0.01**
Сообщение об эмпирической закономерности причинно-следственного типа	0.47±0.01	0.27±0.02	0.34±0.01	0.45±0.01	0.40±0.02	**0.42±0.01**
Классификация: представление результатов классификации данных опыта	0.20±0.04	0.08±0.01	0.12±0.03	0.49±0.11	0.12±0.03	**0.20±0.05**

Оценка качества выявления ментальных действий производилась также на размеченном корпусе. При решении этой задачи подразумевалась, что класс ментальных действий, к которому относятся классифицируемые фрагменты, уже известен, и необходимо лишь выбрать действие в рамках этой группы. Для оценки использовался подход, аналогичный примененному в прошлом эксперименте. Полученные оценки качества для метода, основанного на лесе деревьев решений (RF) и на нейронной сети (NN_2), представлены в Табл. 3.

Табл. 3. Оценки качества выявления ментальных действий (начало)

	RF			**NN_2**		
Ментальное действие	P	R	F_1	P	R	F_1
1.1. Систематизированное описание свойств нового объекта	0.61±0.01	0.95±0.02	0.74±0.01	0.88±0.01	0.92±0.02	0.90±0.01
1.2. Описание комплекса наиболее важных дифференциальных признаков нового объекта	0.63±0.01	0.12±0.01	0.20±0.02	0.82±0.03	0.70±0.01	0.75±0.01
1.3. Определение места нового объекта в системе известных явлений	0.67±0.01	0.15±0.01	0.23±0.01	0.78±0.01	0.77±0.01	0.77±0.01
2.1. Экспликация понятия, передача информации о существенных свойствах и связях предмета	0.54±0.02	0.55±0.03	0.54±0.03	0.82±0.01	0.81±0.01	0.81±0.01
2.2. Формулировка основания деления или ряда делений	0.00±00	0.00±00	0.00±00	0.81±0.09	0.56±0.01	0.67±0.04
2.3. Перечисление выделенных классов	0.49±0.01	0.75±0.01	0.59±0.01	0.78±0.02	0.89±0.06	0.83±0.02
2.4. Диагноз выделяемых классов	0.38±0.02	0.13±0.01	0.20±0.01	0.79±0.03	0.70±0.03	0.74±0.02
3.1. Фиксация данных опыта в разных исследуемых условиях	0.49±0.01	0.93±0.02	0.64±0.01	0.81±0.01	0.86±0.01	0.84±0.01
3.3. Представление установленной причинно-следственной зависимости в ходе ранжирования опытных данных	0.11±0.07	0.01±00	0.01±0.02	0.76±0.05	0.66±0.03	0.71±0.02
3.4. Утверждение о наличии эмпирической закономерности, основанное на статистической обработке данных опыта	0.62±0.01	0.23±0.03	0.34±0.03	0.73±0.03	0.82±0.07	0.77±0.02

Табл. 3. Оценки качества выявления ментальных действий (продолжение)

	RF			**NN_2**		
Ментальное действие	P	R	F_1	P	R	F_1
3.5. Нахождение условий, при которых изменение явления дает наилучший в практическом отношении результат	0.40±0.03	0.11±0.02	0.18±0.02	0.77±0.01	0.75±0.01	0.76±0.01
3.6. Гипотетическое объяснение установленной закономерности более общей закономерностью	0.25±0.10	0.04±0.01	0.07±0.03	0.76±0.01	0.69±0.05	0.73±0.03
4.1. Анализ и интерпретация наиболее влиятельных теорий, формулирование их понятийных систем	0.48±0.01	0.82±0.03	0.61±0.01	0.81±0.01	0.80±0.01	0.81±0.01
4.2. Анализ линии развития научного знания как средство обоснования актуальности изучаемой проблематики	0.44±0.01	0.27±0.03	0.33±0.02	0.72±0.05	0.78±0.01	0.75±0.02
4.3. Определение перспективных направлений разработки проблемы	0.57±0.04	0.08±0.01	0.15±0.02	0.73±0.02	0.70±0.02	0.71±0.01
4.4. Ознакомление с наиболее важными результатами новых исследований и др.	0.51±0.02	0.28±0.01	0.36±0.01	0.75±0.01	0.72±0.05	0.73±0.03
5.1. Формулировка принципов новой теории в форме утверждений о законах, отражающих различные отношения и связи объекта	0.67±0.01	0.16±001	0.26±0.01	0.84±0.04	0.66±0.01	0.74±0.02
5.2. Изложение базовой гипотезы теории	0.62±0.01	0.12±0.01	0.20±0.01	0.85±0.01	0.69±0.07	0.76±0.05
5.3. Обоснование значимости применяемого подхода и демонстрация его преимуществ перед другими теориями	0.68±0.01	0.97±0.01	0.80±0.01	0.87±0.01	0.97±0.01	0.92±0.01
7.1. Формулирование и конкретизация проверяемой гипотезы, определение сути проверяемого предположения	0.42±0.09	0.01±0.01	0.03±0.01	0.67±0.10	0.57±0.07	0.61±0.09
7.2. Описание методики эксперимента	0.46±0.03	0.44±0.02	0.45±0.02	0.73±0.03	0.76±0.03	0.75±0.03
7.3. Описание эксперимента, его условий, факторов, которые могут оказать влияние на его результаты, и самих результатов	0.45±0.02	0.40±0.02	0.42±0.02	0.70±0.05	0.74±0.04	0.72±0.05
7.4. Анализ опытных данных	0.49±0.01	0.81±0.03	0.61±0.01	0.80±0.04	0.78±0.01	0.79±0.02
7.5. Интерпретация опытных данных в форме ответов на вопросы	0.38±0.01	0.19±0.01	0.26±0.01	0.66±0.02	0.67±0.08	0.67±0.04
7.6. Сопоставление результатов теоретической и эмпирической научной деятельности	0.27±0.06	0.04±0.01	0.07±0.01	0.58±0.09	0.55±0.10	0.57±0.09
7.7. Подтверждение или отвержение исходной гипотезы на основе интерпретации опытных данных	0.50±0.03	0.05±0.02	0.10±0.02	0.59±0.10	0.57±0.09	0.58±0.10

Из Табл. 3 видно, что метод на основе сверточной нейронной сети позволяет добиться более высоких оценок качества выявления ментальных действий, чем случайный лес решающих деревьев.

Более высокое качество, полученное с помощью методов на основе нейронных сетей для выявления классов ментальных действий и собственно ментальных действий, скорее всего, связано с тем, что при свертке признаков аргументов предикатных слов в соответствии с формулой (2) теряется часть информации, значимой для классификации.

Заключение

В работе описан размеченный корпус научных публикаций на русском языке, предназначенный для обучения методов выявления ментальных действий, предложены методы автоматического выявления ментальных действий и их классов и произведена их экспериментальная оценка. В Части 2 статьи будет предложен экспериментальный корпус, состоящий из более чем 7000 теоретических и экспериментальных статей на русском языке по различным тематикам, а также представлены результаты экспертной оценки работоспособности предложенных методов на этом корпусе.

Предложенные методы расширяют возможности фактологических и эмпирических исследований в области науковедения, в частности, при изучении вопросов, связанных с нормативностью научного творчества и его спецификой в различных областях научного знания. Разработанные методы также могут лечь в основу программных средств, которые позволили бы выполнять автоматическую классификацию первичных научных публикаций с выделением классов теоретических и экспериментальных статей. В дальнейшем планируется расширить созданный обучающий корпус текстов и использовать разработанные методы для решения перечисленных выше прикладных задач.

Литература

1. Философия. Под ред. Ю.А.Харина. Минск: ТетраСистемс. 2000.
2. Кислов Б.А. О специфике научного метода // Известия ИГЭА. 2004. № 3. С.86-89.
3. Шнякина Н.Ю. Ситуация познания запаха в языковых структурах (субъект – объект – познавательное действие) // Вестник ИГЛУ. 2013. №4 (25). С.121-128.
4. Галич Г.Г. Когнитивные стратегии и языковые структуры. Омск: Изд-во Ом. гос. ун-та, 2011.
5. Червоный А.М. Опыт реконструкции фразеосемантического поля ментальных действий человека на материале французского языка) // Вестник Таганрогского института имени А.П. Чехова. 2016. №2 С.187-193.
6. Баженова Е.А. Стилистико-речевая организация научного текста // Стил. Београд, 2003. № 2. С. 129–141.
7. Баженова Е.А. Прагматические единицы научного текста // http://philologicalstudies.org/dokumenti/2007/vol2/20.pdf
8. Котюрова М.П., Баженова Е.А. Культура научной речи: текст и его редактирование. 2-е изд. М.: Флинта: Наука. 2008.
9. Салимовский В.А. Жанры речи в функционально-стилистическом освещение: Русский научный академический текст. Дисс. … докт.филол.наук. Екатеринбург. 2002.
10. Teufel S., Carletta J., Moens M. An annotation scheme for discourse-level argumentation in research articles //Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. – Association for Computational Linguistics, 1999. P. 110-117.
11. Kirschner C., Eckle-Kohler J., Gurevych I. Linking the thoughts: Analysis of argumentation structures in scientific publications //Proceedings of the 2nd Workshop on Argumentation Mining. 2015. С. 1-11.
12. Liakata M. et al. Corpora for the Conceptualisation and Zoning of Scientific Papers //LREC. – 2010.
13. Guo Y., Reichart R., Korhonen A. Improved Information Structure Analysis of Scientific Documents Through Discourse and Lexical Constraints //HLT-NAACL. 2013. С. 928-937.
14. Liakata M. et al. Automatic recognition of conceptualization zones in scientific articles and two life science applications //Bioinformatics. 2012. Vol. 28. No. 7. P. 991-1000.
15. Liu H. Automatic Argumentative-Zoning Using Word2vec //arXiv preprint arXiv:1703.10152. 2017.
16. Osipov G. et al. Relational-situational method for intelligent search and analysis of scientific publications //Proceedings of the Integrating IR Technologies for Professional Search Workshop. 2013. С. 57-64.
17. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching Word Vectors with Subword Information, arXiv:1607.04606, 2016, unpublished.
18. Осипов Г.С., Девяткин Д.А., Кузнецова Ю.М, Швец А.В. Возможности интеллектуального анализа научных текстов на основе построения когнитивной модели научного текста // Искусственный интеллект и принятие решений (в печати).
19. Chung J. et al. Empirical evaluation of gated recurrent neural networks on sequence modeling //arXiv preprint arXiv:1412.3555. 2014.
20. Bridle J. S. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition //Neurocomputing. – Springer Berlin Heidelberg, 1990. С. 227-236.
21. Zhou Z. H. Ensemble methods: foundations and algorithms. CRC press, 2012.

22. Okazaki N. CRFsuite: a fast implementation of conditional random fields (CRFs). 2007.
23. Breiman L. Random forests //Machine learning. 2001. Т. 45. №. 1. С. 5-32.
24. Pedregosa F. et al. Scikit-learn: Machine learning in Python //Journal of Machine Learning Research. 2011. Т. 12. №. Oct. С. 2825-2830.
25. Kohavi R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection //Ijcai. 1995. Vol. 14(2). p. 1137-1145.
26. Flach P. Machine learning: the art and science of algorithms that make sense of data. – Cambridge University Press. 2012.

Девяткин Дмитрий Алексеевич. Младший научный сотрудник лаборатории «Интеллектуальные технологии и системы» ИСА ФИЦ ИУ РАН. Количество печатных работ: 21. E-mail: devyatkin@isa.ru

Кузнецова Юлия Михайловна. Кандидат психологических наук, старший научный сотрудник ИСА ФИЦ ИУ РАН. Количество печатных работ: 90, в том числе 2 монографии. E-mail: kuzjum@yandex.ru

Чудова Наталья Владимировна. Кандидат психологических наук, старший научный сотрудник ИСА ФИЦ ИУ РАН. Количество печатных работ: более 95, в том числе 2 монографии. E-mail: nchudova@gmail.com

Methods for mental operations detection in scientific publications

D. Devyatkin, Yu. Kuznetcova, N. Chudova

ISA FRC CSC RAS, Moscow, Russia

The paper presents results of a study of methods for a mental structure detection in scientific text. Initially, we consider the concepts of mental operations and mental structure of a scientific text. The reasons for choosing the taxonomy for mental operations used in our approach are also presented. We set the problem of mental operations detection in texts and propose the methods for solving it. An empirical study on an annotated dataset and on a corpus of more than 7k scientific articles showed the applicability of the proposed methods for detection a mental structure of a scientific text. Informative features of mental structures of texts from different research areas were also detected.

Keywords: scientific text, mental operations, representation for mental operations, mental structure of a scientific text, relational-situational analysis.

DOI 10.14357/20718594180203

References

1. Filosofiya [Philosophy]/ Ed. Yu.Kharin. Minsk.TetraSystems. 2000.
2. Kislov B. O specifike nauchnogo metoda [Specifics of the scientific method]// ISEA News. 2004. № 3. p.86-89.
3. Shnyakina N. Situaciya poznaniya zapaha v yazykovyh strukturah (subekt-obekt poznavatelnoe dejstvie) [The situation of recognition of the smell in language structures] (subject-object-cognitive action) // ISLU Bulletin. 2013. №4 (25). p.121-128.
4. Galich G. Kognitivnye strategii i yazykovye struktury [Cognitive strategies and language structures]. Omsk, 2011.
5. Chervonny A. Opyt rekonstrukcii frazeosemanticheskogo polya mentalnyh dejstvij cheloveka na materiale francuzskogo yazyka [Experience in reconstructing the phraseosemantic field of mental actions of a person on the material of the French language] // Bulletin of Taganrog institute after A. Chekhov. 2016(2) p.187-193.
6. Bazhenova E. Stilistiko-rechevaya organizaciya nauchnogo teksta [Stylistic and verbal organization of scientific text] // Beograd, 2003(2). p. 129–141.
7. Bazhenova E. Pragmaticheskie edinicy nauchnogo teksta [Pragmatic units of scientific text] // http://philologicalstudies.org/dokumenti/2007/vol2/20.pdf.
8. Kotyurova M.P., Bazhenova E. A. Kultura nauchnoj rechi. Tekst i ego redaktirovanie [Culture of scientific speech. Text and editing], M. Flinta. 2008.
9. Salimovskij V.A. Zhanry rechi v funkcionalno-stilisticheskom osveshhenii russkij akademicheskij tekst [Speech genres in functional-stylistic coverage of Russian academic text]. Diss dok filol nauk [Dr. Sci theses]. Ekaterinburg. 2002.
10. Teufel S., Carletta J., Moens M. An annotation scheme for discourse-level argumentation in research articles //Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. – Association for Computational Linguistics, 1999. – P. 110-117.
11. Kirschner C., Eckle-Kohler J., Gurevych I. Linking the thoughts: Analysis of argumentation structures in scientific publications //Proceedings of the 2nd Workshop on Argumentation Mining. – 2015. – p. 1-11.
12. Liakata M. et al. Corpora for the Conceptualisation and Zoning of Scientific Papers //LREC. – 2010.
13. Guo Y., Reichart R., Korhonen A. Improved Information Structure Analysis of Scientific Documents Through Discourse and Lexical Constraints //HLT-NAACL. – 2013. – С. 928-937.

14. Liakata M. et al. Automatic recognition of conceptualization zones in scientific articles and two life science applications //Bioinformatics. – 2012. – Vol. 28. – No. 7. – P. 991-1000.
15. Liu H. Automatic Argumentative-Zoning Using Word2vec //arXiv preprint arXiv:1703.10152. – 2017.
16. Osipov G. et al. Relational-situational method for intelligent search and analysis of scientific publications //Proceedings of the Integrating IR Technologies for Professional Search Workshop. – 2013. – p. 57-64.
17. P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching Word Vectors with Subword Information”, arXiv:1607.04606, 2016, [unpublished].
18. G. Osipov, D. Devyatkin, Yu. Kuznetcova, A. Shvets Vozmozhnosti intellektualnogo analiza nauchnyh tekstov na osnove postroeniya kognitivnoj modeli nauchnogo teksta [possibilities of intellectual analysis of scientific texts by construction of their cognitive models] // Artificial Intelligence and Decision Making [in print].
19. Chung J. et al. Empirical evaluation of gated recurrent neural networks on sequence modeling //arXiv preprint arXiv:1412.3555. – 2014.
20. Bridle J. S. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition //Neurocomputing. – Springer Berlin Heidelberg, 1990. – p. 227-236.
21. Zhou Z. H. Ensemble methods: foundations and algorithms. – CRC press, 2012.
22. Okazaki N. CRFsuite: a fast implementation of conditional random fields (CRFs). – 2007.
23. Breiman L. Random forests //Machine learning. – 2001. – Vol. 45. – №. 1. – p. 5-32.
24. Pedregosa F. et al. Scikit-learn: Machine learning in Python //Journal of Machine Learning Research. – 2011. – Vol. 12. – No. Oct. – p. 2825-2830.
25. Kohavi R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection //Ijcai. – 1995. – Vol. 14(2). – p. 1137-1145.
26. Flach P. Machine learning: the art and science of algorithms that make sense of data. – Cambridge University Press. 2012.

Devyatkin Dmitry A. Researcher at “Intelligent technologies and systems” laboratory within ISA FRC CSC RAS, Moscow, pr. 60-letiya Oktyabrya, 9. Authored 21 scientific papers. E-mail: devyatkin@isa.ru

Kuznetsova Yulia M. PhD, Senior researcher at ISA FRC CSC RAS Moscow, pr. 60-letiya Oktyabrya, 9. Authored 90 scientific papers, 2 monographs. E-mail: kuzjum@yandex.ru

Chudova Natalia V. PhD, Senior researcher at ISA FRC CSC RAS Moscow, pr. 60-letiya Oktyabrya, 9. Authored 95 scientific papers, 2 monographs. E-mail: nchudova@gmail.com