Разработка методов анализа сценариев поведения (на материале инструктивных интернет-текстов)*

В. А. Мишланов¹, А. А. Чуганская², И. В. Смирнов²,
М. И. Суворова², И. А. Курузов³

¹ Пермский государственный национальный исследовательский университет,
Российская Федерация, 614068, Пермь, ул. Букирева, 15

² Институт проблем искусственного интеллекта ФИЦ «Информатика и управление» РАН,
Российская Федерация, 117312, Москва, пр. 60-летия Октября, 9

³ Московский физико-технический институт (национальный исследовательский университет),
Российская Федерация, 141701, Московская обл., Долгопрудный, Институтский пер., 9

Для цитирования: Мишланов, В. А., Чуганская, А. А., Смирнов, И. В., Суворова, М. И., Курузов, И. А. (2020). Разработка методов анализа сценариев поведения (на материале инструктивных интернет-текстов). *Медиалингвистика*, 7 (1), 16–28. <https://doi.org/10.21638/spbu22.2020.102>

Рассматриваются вопросы моделирования сценариев поведения и методов их извлечения из текстов, циркулирующих в социальных медиа. Авторы исходят из базового в современной когнитивистике положения, согласно которому интенциональный аспект поведения человека отражается главным образом в его речевой деятельности, что дает основание «сценарное» воплощение деятельности, протекающей в разных ситуациях, извлекать из текстов, порождаемых в связи с этой деятельностью. Выбор в качестве материала исследования инструктивных текстов определяется тем, что мотивирующие их порождение коммуникативные интенции, жанровая форма и содержание, наиболее эксплицитно воплощают заключенную в них программу поведения человека. Предлагаются варианты методик автоматического построения сценариев поведения. Представлена разработанная в виде фрейма модель сценария, которая может быть использована в качестве основы автоматической системы сценарного анализа текстов инструкций и дальнейшего сопровождения пользователя данной системы. Описаны результаты эксперимента по автоматическому выявлению сценариев из текстов на естественном языке. Процесс извлечения сценария включает этап сегментации текста, в результате которой каждому выделенному сегменту ставится в соответствие отдельный шаг сценария. Дается описание процедур: разбиение текста на смысловые единицы, выбор наиболее важных фрагментов — центров, выражающих (лексическим значением и формой) тот или иной вид побудительной модальности, присоединение к центрам контекста. В психологическом плане это означает, что на данном этапе извлечения сценарной информации в инструктивном тексте выделяются сегменты, содержащие самостоятельные цели и описывающие действия, приводящие, по мнению автора текста, к их достижению. Реконструкция сценария мыслится как «сборка» целей, акторов, действий и условий, идентифицированных по этим функциям при автоматическом лингвистическом анализе выделенных сегментов. Это дало содержательно

* Работа выполнена при частичной поддержке РФФИ (гранты № 17-07-00651 «Разработка моделей и методов конструирования сценариев поведения на основе анализа текстов» и № 18-29-22027 «Персональные когнитивные ассистенты, сопровождающие деятельность человека в информационном пространстве»).

обоснованный результат разбиения текста на сегменты, которые в последующем классифицируются с достаточно высоким качеством по F-мере.

Ключевые слова: сценарии поведения, инструктивные тексты, социальные медиа, модель сценария.

Постановка проблемы

Развитие средств массовой коммуникации в современном информационном обществе обусловило весьма значительные изменения в психологических и лингвистических механизмах общения [Баранов 2001]. Степень этих изменений в новой коммуникативной среде достигает значительных масштабов, позволяющих говорить о формировании новых моделей речевого взаимодействия. Для выявления сущности происходящих изменений в различных сферах коммуникации стоит обратиться к анализу сценариев речевого поведения человека в определенных ситуациях.

Представляя в словесной форме сценарий поведения, человек опирается на приобретенный им коммуникативный опыт — знание принятых в социуме стереотипов речевого взаимодействия [Кузнецова 2019]. Анализ сценариев дает ценный материал для решения значимой психологической и лингвистической задачи объективации знания в сфере социальных отношений [Кузнецова 2018]. В частности, сценарные тексты позволяют получить представления о принятых нормах речевого поведения в различных ситуациях социального взаимодействия [Кузнецова и др. 2019a]. В исследовательском плане особо значимыми, на наш взгляд, являются обобщенные сценарии, представленные в «виртуальном» общении: они дают богатый материал для проведения исследований в интересах различных когнитивных дисциплин [Кузнецова, Чудова 2008].

Целью статьи является описание особенностей анализа инструктивных текстов (на примере текстов, связанных с ситуацией покупки автомобиля) [Смирнов и др. 2018], направленного на построение модели «сценарного» речевого поведения и создание оптимальных методов извлечения сценариев и их компонентов из текстового материала.

История вопроса

В когнитивных науках сценарий, наряду с понятиями фрейма, ситуативной модели и другими [Волосухина 2010], рассматривается в качестве научного конструкта, с помощью которого можно осуществить моделирование ментальных состояний субъекта [Демьянков 1994].

Одним из наиболее известных направлений в исследовании сценарных моделей поведения является фреймовый подход, опирающийся на предложенное М. Минским понятие «фрей» [Минский 1979] как структуры данных, предназначенной для описания стереотипной ситуации.

Наряду с фреймами, содержащими декларативные (дескриптивные) знания, было предложено понятие сценария, хранящего знания о динамических явлениях, которые поэтому могут быть представлены как ряд сменяющих друг друга состояний.

Ч. Филмор рассматривает сценарий как фиксированную в языке когнитивную структуру, которая включает варианты привычных межличностных интеракций, реализуемых согласно нормам той или иной культуры, а также институциональных структур, аккумулирующих опыт профессионального взаимодействия (см. [Кузнецова 2018]). Р. Шенк и Р. Абельсон в сходном значении используют термин «скрипт», понимая под ним предопределенную последовательность стереотипных действий, характерных для общеизвестной ситуации (см. [Полатовская 2013]).

В качестве особого типа выделяются коммуникативные сценарии, или сценарии речевого поведения. По определению В. И. Шляхова, они представляют собой статико-динамические структуры, включающие несколько речевых действий, которые связаны между собой иерархическими отношениями. Автор подчеркивает конвенциональную ценность коммуникативных сценариев: «Социум выработал правила и схемы речевого поведения и ожидает от индивида определенных действий в определенных обстоятельствах» [Шляхов 2007: 26].

Отметим в этой связи, что лингвистика в союзе когнитивных наук приобретает все возрастающую роль не только потому, что она представляется «филиалом когнитивной психологии» и использует «арсенал переработки языковой информации для построения моделей, имитирующих внешние проявления человеческого поведения при решении интеллектуальных задач» [Демьянков 1994: 18], но и в силу того особого обстоятельства, которое в человеческой деятельности занимает речевая коммуникация. В сущности, текст как воплощение (объективация) деятельности, реализуемой в дискурсе, пока едва ли не единственный источник сценарного моделирования поведения, поэтому оптимальным материалом для исследования сценарного поведения становятся не просто тексты на естественном языке, а такие речевые произведения, которые по своей иллокутивной природе и являются самой деятельностью (тексты политико-административной, педагогической, массмедийной, рекламной и тому подобной деятельности).

Неслучайно сценарный подход привлек большое внимание в рамках нарративного направления коммуникативистики и гуманитарного знания в целом (Э. Эббот, Л. Гриффин и др.) [Брокмейер, Харре 2000]. Методологической особенностью этого направления стало стремление исследовать социальные феномены как представленные в тексте (нарративе) последовательности событий [Abbot 1992], которые только так, в текстовом воплощении, отражающем точку зрения рассказчика, и могут стать предметом научного осмысления [Брокмейер, Харре 2000]. Из этого следует, что к «текстоцентрическим» наукам относится не только филология, для которой «исходной реальностью» является «текст во всей совокупности своих внутренних аспектов и внешних связей» [Аверинцев 1979: 372], но и все другие гуманитарные дисциплины.

В нарративном направлении стоит выделить подход Х. Олкера, который предложил использовать для аналитического описания международных событий инструментарий, ведущий свое происхождение от работ отечественного филолога В. Я. Проппа [Чуганская 2019]. Исследуя тексты русских волшебных сказок, В. Я. Пропп выделяет устойчивую структуру: определенный круг основных действующих лиц и определенное множество (более 30) их основных функций [Пропп 1998]. По мнению Х. Олкера, аналогичный каркас обнаруживается в структуре сценария тех или иных событий общественной жизни [Олкер 1987: 33].

В лингвистических работах когнитивистского направления понятие «сценарий» включает такие признаки содержания текста, как типизированность, наличие определенного набора участников описываемого действия, динамичность, связанность с тем или иным видом социальной практики [Кузнецова и др. 20196]. Важной характеристикой является также иерархичность структуры сценария, в которой компоненты верхнего уровня отражают устойчивые признаки, обязательные с точки зрения содержания сценария, а элементы низших уровней наполняются в зависимости от конкретной ситуации [Кузнецова 2018].

Будучи регулятором коммуникативного поведения, сценарий реализуется в конкретной проблемной ситуации, разрешение которой требует формирования некоего плана действий для адаптации имеющихся сценариев к конкретной ситуации [Кузнецова и др. 20196]. В отличие от сценария, план связан с конкретной ситуацией. При формировании плана существенно важным оказывается влияние картины мира субъекта, частью которой является сценарий [Осипов и др. 2017].

Связь сценариев с естественным языком определяет такое их качество, как культуроспецифичность. Используемые языковыми коллективами в разных ситуациях общения коммуникативные стратегии, по мнению А. Вежбицкой, представляют собой речевое выражение скрытой (неписаной) системы культурных правил, или культурных сценариев, отражающих этнические установки и нормы поведения [Вежбицкая 1999].

В настоящей работе предпринята попытка использования сценарного подхода к анализу инструктивных текстов, содержание и коммуникативный смысл (интенции) которых в наибольшей степени отвечают базовому понятию сценария. Такие тексты содержат, как правило, прямые наименования основных действий, состояний, признаков ситуации, в которой инструктируемый (коллективный адресат) мыслится автором текста как исполнитель определенной роли. Коммуникативная цель текста-инструкции заключается в том, чтобы донести до адресата содержание тех стандартов действий, соблюдение которых признается социумом существенно важным для успешного осуществления различных жизненных практик. Четкие и недвусмысленные формулировки, наличие эксплицитной мотивирующей составляющей, усиливающей побудительную модальность текста, важны с точки зрения минимизации поведенческой вариативности, что в итоге обеспечивает деятельность, необходимую для поддержания социума. Человек, не имеющий доступа к инструктивным текстам, может испытывать сложности в регламентированных ситуациях, в которых владеющий сценариями обычно решает задачи автоматически [Кузнецова и др. 2019а].

Представляется целесообразным использовать сценарный подход при изучении потребительского поведения (например, покупки автомобиля). В основе подхода лежит лингвистический (интенциональный, лексико-семантический и грамматический) анализ инструктивных текстов, который может быть осуществлен в трех аспектах:

- функциональном (коммуникативном), предусматривающем выявление коммуникативной интенции, предметной цели, композиции текстовых фрагментов;
- структурном, состоящем в выделении действующих лиц и операндов: предметов-целей и предметов-условий;

- процессуальном, имеющем целью описание «шагов» и «развилки», или точек выбора (анализ в этом аспекте может быть проведен с ориентацией на идеи В. Я. Проппа и Х. Олкера [Чуганская 2019]).

Описание методики исследования

В качестве материала анализа используются извлеченные из Интернета инструкции по покупке автомобиля. Корпус насчитывает 100 уникальных текстов (объемом 147 445 слов), содержащих поэтапное описание процесса приобретения автомобиля. Тексты были размечены вручную тремя экспертами, результаты разметки использовались при создании общей модели сценария.

В контексте задач автоматического анализа текстов (в нашем случае идентификации фрагментов инструктивных текстов как определенных шагов сценария) особое значение приобретает метод реляционно-ситуационного анализа [Смирнов и др. 2018], основной задачей которого является «выявление значений синтаксисом и семантических связей между ними» [Осипов и др. 2008: 5]. В основу метода положен анализ глаголов и других предикатных слов, определяющих сочетаемость с синтаксемами и структуру предложения в целом. Такой анализ важен и в аспекте изучения строения деятельности, в частности он позволяет адекватно определить целевой компонент того или иного действия в сценарии.

В решении задачи автоматического построения сценария можно выделить два этапа: анализ структуры инструктивного текста в означенных выше трех аспектах и синтез сценария на основе полученной сценарной информации. В нашей работе рассматривается первый этап, предполагающий сегментацию текста на фрагменты, каждый из которых посвящен достижению какой-либо одной цели (анализ структурного аспекта). Внутри каждого фрагмента автоматический анализ выявляет фрагменты, описывающие условия, в которых дана цель, их вариации и операции, отвечающие обсуждаемым в тексте условиям (анализ процессуального аспекта). Фрагменты, содержащие коммуникативные цели автора инструктивного текста, при ручной разметке были отделены от основного корпуса, в котором содержатся фрагменты, описывающие цели инструктируемого читателя (в нашем случае — потенциального покупателя автомобиля), так что на данном этапе исследования для функционального анализа методы автоматического извлечения сценарной информации не разрабатывались. В рамках психологической структуры сценарного поведения каждый шаг сценария, представленный выделяемым фрагментом текста, может быть определен как отдельное действие в составе деятельности по покупке автомобиля. Таким образом, метод автоматического анализа структурного и процессуального аспектов инструктивного текста представляет собой автоматизированную процедуру выделения единиц деятельности — действий и операционального состава деятельности [Леонтьев 1974].

Рассмотрим подробнее общую модель сценария, принятую в исследовании. Обобщенная структура данных, предложенная авторами статьи, представлена на рисунке. В основу ее лег метод представления знаний из области искусственного интеллекта, основывающийся на понятии фрейма [Минский 1979]. Данный фрейм-сценарий представляет собой вложенную структуру, где каждый шаг является отдельным фреймом.

Рис. Общая схема фрейма-сценария

На основе анализа инструктивных текстов были выделены три вида сценариев:

1. *Простое действие*, представляющее собой один конкретный сценарий, который необходимо выполнить (например, поиск объявлений о продаже подержанных авто или выбор функционала автомобиля).
2. *Ветвления*, или множество сценариев, включающих альтернативные шаги, необязательно в определенной последовательности. Примерами ветвлений являются сценарии, предусматривающие покупку иномарки или отечественной машины, в кредит или за наличные средства, нового или подержанного автомобиля.
3. *Цепочка шагов*, или множество сценариев, которые необходимо выполнить в строго заданной последовательности (выбор конкретной модели, осмотр автомобиля в салоне, оформление покупки). Этот вид сценария допускает возвращение на предыдущие шаги (если при осмотре автомобиля выявились серьезные недостатки, можно выбрать новый вариант).

К другим компонентам фрейма относятся активные участники сценария (покупатель, продавец, страховой агент, представитель ГИБДД) и операнды: автомобиль, его внешние свойства (лакокрасочное покрытие, кузов и др.), договор купли-продажи, а также цель, т.е. желаемая ситуация, которая будет достигнута после выполнения действий сценария (описывается хранящимися в рабочей памяти «фактами о мире»), и условия, в которых событие происходит.

Для начала выполнения любого сценария необходимо осуществить определенные предусловия (также в рабочей памяти представленные фактами о мире): принять решения (купить новый автомобиль), найти необходимую информацию, оценить ресурсы (покупатель имеет информацию о марках автомобиля, официальных представителях фирмы, располагает средствами и временем на покупку автомобиля и др.).

Первым этапом обработки текстов инструкций, как уже было сказано, является сегментация, в результате которой исходный текст разделяется на фрагменты, соответствующие отдельным шагам сценария.

Сегментацию можно разделить условно на четыре стадии. На первой осуществляется разбиение текста на цельные смысловые единицы, под которыми понимаются предложения и списки, т. е. однородные компоненты предложения, выделенные в тексте таким образом, что каждый из них находится на отдельной строке.

Вторая стадия состоит в том, чтобы среди смысловых единиц выделить элемент «рекомендуемое действие» — центры будущих сегментов, к которым мы относим предикативные основы высказываний: глаголы или глагольные сочетания, выражающие (лексическим значением и формой) тот или иной вид побудительной модальности — совет, рекомендацию, призыв и т. п. Такими центрами являются, например, словоформы 2 л. мн. ч. повелительного наклонения глаголов (*определи-тесь, сделайте*), модальные наречия или безличные глаголы с инфинитивом (*можно сделать, нужно выбрать, следует оформить* и т. п.) и др.

Несмотря на то что смысловые центры сегментов (предикаты побудительной модальности) содержат важнейшую информацию о сценарном действии, использовать для создания сценария только эти данные недостаточно. Это связано с тем, что, как показали эксперименты, при рассмотрении множества таких центров без учета контекста соответствующие им векторные представления оказываются слабо разделимыми, при этом снижается качество классификации сегментов.

На третьей стадии каждому элементу, полученному при разбиении текста, необходимо поставить в соответствие некоторое векторное представление. В рамках исследования сценариев в инструктивных текстах каждому слову мы ставили в соответствие вектор, используя готовые модели word2vec из RusVectores [Kutuzov, Kuzmenko 2017], а вектор для элемента текста находили как среднее арифметическое векторов для слов, входящих в этот элемент, кроме стоп-слов. Подобное векторное представление позволяет представить большинство элементов как достаточно короткие тексты. Как было показано в более ранних исследованиях [Mikolov et al. 2013], данный методический прием оправдан.

Поставив в соответствие каждому элементу текста вектор, мы можем в дальнейших рассуждениях отождествлять понятия «элемент текста» и «вектор». Все элементы перед первым центром мы относим к первому центру, все элементы после последнего — к последнему. Остальные элементы находятся между двумя центрами, и для всех пар соседних центров мы находим разбиение предложений между ними на два непересекающихся множества: контекст первого центра и контекст второго центра. Естественно предположить, что элементы различных контекстов не должны чередоваться. Поэтому мы ставим дополнительное условие: разбиение должно быть таким, что все элементы второго множества в тексте находятся после первого множества.

Формализуем задачу нахождения этого разбиения. Пусть расстояние от контекста до центра — это сумма расстояний от предложений контекста до центра. Тогда найдем разбиение, минимизирующее расстояние между соответствующими центрами и контекстами. Эта задача решается путем перебора всевозможных разбиений за линейное время от количества предложений между центрами. В качестве расстояния от предложения до центра мы использовали евклидово расстояние между соответствующими векторными представлениями.

Последней, четвертой, стадией сегментации является объединение фрагментов, которые составляют один и тот же шаг в рамках сценария. Мы получаем векторные представления для сегментов аналогично векторным представлениям для элементов, которые описаны выше. Для каждой пары соседних сегментов было посчитано расстояние между ними и объединены в группу соседствующих сегментов, если расстояние между парами соседних сегментов в этой группе было меньше порога, устанавливаемого экспериментально.

В качестве расстояния мы использовали взвешенную линейную комбинацию из WMdistance и сигмоида от суммы длин сегментов, где WMdistance — Word Mover's Distance, один из множества вариантов измерения семантической близости между двумя текстами [Kusner et al. 2015], а сигмоид понимается как монотонно возрастающая ограниченная функция. Использование WMdistance обусловлено тем, что для сегментов, которые являются относительно большими текстами, он показал лучшие результаты, чем остальные метрики. Мы предполагаем также, что короткие сегменты, скорее всего, не представляют сами по себе большой ценности, поэтому мы используем сигмоид, который поощряет объединение коротких и штрафует объединение длинных сегментов, причем штраф практически не различается для длинных и сверхдлинных текстов.

Полученные сегменты используются для формирования нового сценария или улучшения уже существующего. Во втором случае имеется уже схема шагов, а для каждого шага — некоторая коллекция соответствующих сегментов, и может быть поставлена задача интерпретировать шаги как классы, по которым необходимо распределить наши объекты (это стандартная задача классификации и возможности ее решения будут обсуждены ниже).

Результаты

Перейдем к анализу результатов решения сформулированных выше задач — выделения сегментов, воплощающих определенный шаг сценария, их классификации.

При автоматическом анализе сценариев в инструктивных текстах важно учитывать, в какой мере может быть формализовано разграничение соседних шагов одного текста-сценария и насколько достижимо отождествление фрагментов разных текстов как представляющих один и тот же шаг сценария.

Мы предположили, что множество шагов в векторном пространстве имеет довольно простую структуру, а именно: каждый шаг имеет свой центр (вектор), и разбиение Вороного [Aurenhammer 1991], построенное на этих центрах, задает корректное разделение векторного пространства на шаги.

С целью проверки этого предположения проведен следующий эксперимент: вручную выбранные и размеченные сегменты были отображены в векторное пространство, для каждого шага оценен соответствующий центр как среднее арифметическое векторов для выбранных сегментов, для каждого сегмента определен шаг по ближайшему центру и проведено сравнение полученной разметки с исходной.

На основе анализа инструктивных текстов было выделено 12 шагов, для каждого из которых найдено от трех до девяти текстовых сегментов:

1. Ваши деньги.
2. Цены.
3. Объявления.
4. Телефонный разговор.
5. Документы на машину.
6. Мониторинг сайтов.
7. ДКП.
8. Осмотр.
9. Тест-драйв.
10. Марка и модель машины.
11. Диагностика.
12. Год выпуска.

В наших экспериментах для оценки качества каждого класса мы использовали F-меру с параметром 0.5, т.е. среднее геометрическое полноты (Recall) и его чистоты (Precision) полученного класса, давая последней больший вес [Hastie, Tibshirani, Friedman 2009]. Для суммарной оценки качества мы использовали усредненную F-меру.

При анализе вручную выделенных сегментов из инструктивных текстов были получены неплохие, на наш взгляд, результаты:

- среднее значение F-меры — 89.6,
- минимальное значение F-меры: — 68.2.

Следует, однако, заметить, что некоторые шаги (например, четвертый) распознаются хуже. Это связано с тем, что такие шаги имеют существенное пересечение, так как включают однотипную лексику.

Исследование показало также, что если при автоматическом выделении шагов сценария инструктивных текстов центры оцениваются не на всем множестве размеченных текстов, а только на части, то качество классификации на оставшейся выборке сколько-нибудь существенно не снижается (так, если мы отложим треть данных как тестовую, значение F-меры падает лишь до 80 %).

Добавим, что полученные в ходе эксперимента результаты позволяют считать обоснованным предположение о том, что структура выявляемых в тексте шагов сценария не отличается сложностью. В рамках исследования на вручную выделенных сегментах мы показали, что выделенные экспертами части текста можно классифицировать вполне качественно, имея разметку даже для небольшого количества сегментов.

Одной из целей эксперимента была оценка возможностей решения задачи классификации автоматически выделенных объектов. Сегменты, полученные программой, были промаркированы, в результате чего была получена выборка, включающая 490 фрагментов. В этой выборке отсутствует пятый шаг (мониторинг сайтов), поскольку для него нашлось только два сегмента. Выборка оказалась не вполне сбалансированной, в частности седьмой шаг (осмотр) составляет почти половину выборки (225 сегментов).

Обсудим методы решения задачи автоматической классификации сегментов, начав с рассмотрения «наивного» классификатора, основанного на оценке центров.

Наивная классификация предполагает применение той же модели, которая была использована в предыдущем разделе. Каждому сегменту ставится в соответствие вектор как среднее арифметическое векторов предложений, которые, в свою очередь, есть среднее арифметическое векторов слов. По размеченным данным мы получаем центр каждого класса как среднее арифметическое соответствующих векторов из обучающей выборки. Для новых объектов класс определяется по ближайшему вектору.

Далее рассмотрим логистическую регрессию [Hastie, Tibshirani, Friedman 2009] с L2-регуляризацией, которая показала один из наилучших результатов среди различных методов. Параметр регуляризации подбирался с учетом кросс-валидации по пяти блокам.

Для измерения качества использовалась вышеописанная F-мера. Качество наивной классификации определялось на кросс-валидации по пяти блокам. Для логистической регрессии качество модели измерялось на тестовой выборке, составляющей 15 % от исходной. Были получены следующие результаты среднего значения F-меры:

- наивный классификатор: 52 %;
- логистическая регрессия: 59 %.

Наивная классификация, строящая разбиение Вороного на основе оценки центра, дает результат лучше, чем вышеописанный константный классификатор. Это говорит о том, что структура шагов, рассмотренная для выделенных человеком сегментов, имеет место и для шагов, построенных на программно выделенных сегментах. Отметим, что логистическая регрессия заметно повышает качество на большинстве шагов (в среднем на 7 %).

Выводы

Нами предложена конкретная структура фрейма, которая схематически описывает сценарий. Одним из достоинств данной структуры является ее универсальность. Практическая значимость такой структуры данных состоит в операционализированности, т. е. в возможности быть использованной при анализе сценарного поведения субъекта, в том числе для автоматического извлечения сценария из текстов на естественном языке с использованием технологий искусственного интеллекта.

Кроме того, была предложена методика автоматического решения задачи сегментации инструктивного текста, позволяющая выделить шаги сценария, которые отражают основные параметры действия как подчиненного определенной цели, и классификации полученных сегментов. Исследование показало, что предложенная схема сегментации достаточно эффективна, позволяет получать фрагменты, содержащие необходимую и достаточную информацию об одном шаге сценария. Проведенные эксперименты показали, что множество шагов имеют в векторном представлении довольно простую структуру, и продемонстрировали приемлемое качество классификации полученных сегментов.

Литература

- Аверинцев, С. С. (1979). *Филология. Русский язык: энциклопедия*. Москва: Советская энциклопедия.
- Баранов, А. Н. (2001). *Введение в прикладную лингвистику*. Москва: Эдиториал УРСС.
- Брокмeyer, Й., Харре, Р. (2000). Нарратив: проблемы и обещания одной альтернативной парадигмы. *Вопросы философии*, 3, 29–42.

- Вежбицкая, А. (1999). *Семантические универсалии и описание языков*. Москва: Языки русской культуры.
- Волосухина, Н. В. (2010). К вопросу о трактовке понятий «концепт» и «фрейм» в современной лингвистике. В *Материалы научно-методических чтений ПГЛУ* (с. 41–46). Пятигорск: Изд-во Пятигорского государственного лингвистического университета.
- Демьянков, В. З. (1994). Когнитивная лингвистика как разновидность интерпретирующего подхода. *Вопросы языкознания*, 4, 17–19.
- Кузнецова, Ю. М. (2018). Сценарный подход к анализу текстов. *Труды ИСА РАН*, 1 (68), 31–41.
- Кузнецова, Ю. М. (2019). Социальные сценарии поведения как предмет сетевых обсуждений. В *Категория «социального» в современной педагогике и психологии*. Мат-лы 7-й Всерос. науч.-практ. конф. с дистанц. и междунар. участием (с. 177–183). Ульяновск: Зебра.
- Кузнецова, Ю. М., Осипов, Г. С., Смирнов, И. В., Чудова, Н. В. (2019а). Текст сетевой дискуссии как источник сценарной информации. *Речевые технологии*, 1, 30–44.
- Кузнецова, Ю. М., Суворова, М. И., Чудова, Н. В. (2019б). Сценарий как форма репрезентации события в знаковой картине мира. *Труды ИСА РАН*, 1, 70–82.
- Кузнецова, Ю. М., Чудова, Н. В. (2008). *Психология жителей Интернета*. Москва: Изд-во ЛКИ.
- Леонтьев, А. Н. (1974). *Деятельность. Сознание. Личность*. Москва: Политиздат.
- Минский, М. (1979). *Фреймы для представления знаний*. Москва: Энергия.
- Олкер, Х. Р. (1987). Волшебные сказки, трагедии и способы изложения мировой истории. В *Язык и моделирование социального взаимодействия* (с. 408–440). Москва: Прогресс.
- Осипов, Г. С., Панов, А. И., Чудова, Н. В., Кузнецова, Ю. М. (2017). *Знаковая картина мира субъекта поведения*. Москва: Физматлит.
- Осипов, Г. С., Смирнов, И. В., Тихомиров, И. А. (2008). Реляционно-ситуационный метод поиска и анализа текстов и его приложения. *Искусственный интеллект и принятие решений*, 2, 3–10.
- Полатовская, О. С. (2013). Фрейм-сценарий как тип концептов. *Вестник Иркутского государственного лингвистического университета*, 4 (25), 161–166.
- Пропт, В. Я. (1998). *Морфология (волшебной) сказки: исторические корни волшебной сказки*. Москва: Лабиринт.
- Смирнов, И. В., Шелманов, А. О., Исаков, В. А., Станкевич, М. А. (2018). Открытое извлечение информации из текстов. Ч. I. Постановка задачи и обзор методов. *Искусственный интеллект и принятие решений*, 2, 47–61. DOI: 10.14357/20718594180204.
- Чуганская, А. А. (2019). Сценарный подход в ассистировании экономического поведения покупателя. В *Ломоносовские чтения — 2019. Секция экономических наук. Экономические отношения в условиях цифровой трансформации: сборник тезисов выступлений* (с. 439–442). Москва: Изд-во Московского университета.
- Шляхов, В. И. (2007). Сценарии русского речевого взаимодействия. *Русский язык за рубежом*, 2, 26–34.
- Abbot, A. (1992). From causes to events. Notes on narrative positivism. *Sociological methods & research*, 20 (4), 428–455.
- Aurenhammer, F. (1991). *Voronoi Diagrams — A Survey of a Fundamental Geometric Data Structure*. *ACM Computing Surveys*, 23 (3), 345–405. DOI: 10.1145/116873.116880.
- Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. 2nd ed. Berlin: Springer.
- Kusner, M., Sun, Y., Kolkin, N. I., Weinberger, K. Q. (2015). *From embeddings to document distances*. The 32nd International Conference on Machine Learning. Lille, France (pp. 957–966). Lille: ICML.
- Kutuzov, A., Kuzmenko, E. (2017). WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In D. Ignatov et al. (Eds), *Analysis of Images, Social Networks and Texts*. AIST 2016. Communications in Computer and Information Science. Vol 661 (pp. 155–161). Springer: Cham. DOI 10.1007/978-3-319-52920-2-15. Электронный ресурс <https://rusvectors.org/ru/>.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013). Distributed Representations of Word and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119). Электронный ресурс <https://arxiv.org/abs/1310.4546>.

Статья поступила в редакцию 1 октября 2019 г.;
рекомендована в печать 10 ноября 2019 г.

Контактная информация:

Мишланов Валерий Александрович — д-р филол. наук, проф.; mishl@psu.ru
Чуганская Анфиса Анваровна — канд. психол. наук; anfisa.makh@gmail.com
Смирнов Иван Валентинович — канд. физ.-мат. наук; ivs@isa.ru
Суворова Маргарита Игоревна — аспирант; suvorova@isa.ru
Курузов Илья Алексеевич — аспирант; kuruzov2014@mail.ru

Developing methods for behavior scenario analysis (on the material of instructional texts)

V. A. Mishlanov¹, A. A. Chuganskaya², I. V. Smirnov², M. I. Suvorova², I. A. Kuruzov³

¹ Perm' State University,
15, ul. Bukireva, Perm', 614068, Russian Federation

² Artificial Intelligence Research Institute of Federal Research Center
'Computer Science and Control', Russian Academy of Sciences,
9, pr. 60-letia Oktiabria, Moscow, 117321, Russian Federation

³ Moscow Institute of Physics and Technology (National Research University),
9, Institutskii per., Dolgoprudnyi, Moscow region, 141701, Russian Federation

For citation: Mishlanov, V. A., Chuganskaya, A. A., Smirnov, I. V., Suvorova, M. I., Kuruzov I. A. (2020). Developing methods for behavior scenario analysis (on the material of instructional texts). *Media Linguistics*, 7 (1), 16–28. <https://doi.org/10.21638/spbu22.2020.102> (In Russian)

The article discusses the modeling of behavioral scenarios and methods for their extraction from texts published in social media. The authors proceed from the basic (in modern cognitive science) position that the intentional aspect of a person's behavior is reflected in his speech activity. This gives reason to extract scenarios from the texts generated to carry out some activity. The choice of instructive texts is determined by the fact that their genre, content and communicative intentions of their authors most explicitly represent the program of human behavior. In this article we propose several methods for automatic construction of behavior scenarios, as well as a scenario model in the form of a frame. This model can be used in the system for automatic extraction of scenarios from texts and their further implementation. We describe an experiment on automatic scenario identification in texts and its results. The experiment includes text segmentation after which each segment is associated with a particular step of a scenario. Text segmentation involves dividing a text into semantic units, selecting the most important of them, which express (by their lexical meaning and form) incentive modality, and finding some details to these units. In psychological terms, this means that we extract those segments that contain description of some goals and particular actions that lead to their achievement. Reconstruction of the scenario in this case can be represented as automatic extraction of goals, actors, actions and conditions from selected segments. The result of text segmentation achieves a fairly high F-measure.

Keywords: behavior scenarios, instructional texts, social media, scenario model.

References

- Abbot, A. (1992). From causes to events. Notes on narrative positivism. *Sociological methods & research*, 20 (4), 428–455.
- Aurenhammer, F. (1991). *Voronoi Diagrams — A Survey of a Fundamental Geometric Data Structure*. *ACM Computing Surveys*, 23 (3), 345–405. DOI: 10.1145/116873.116880.
- Averintsev, S. S. (1979). *Philology. Russian language: encyclopedia*. Moscow: Soviet Encyclopedia Publ. (In Russian)
- Baranov, A. N. (2001). *Introduction in applied linguistics*. Moscow: Editorial URSS Publ. (In Russian)

- Brokmeier, I., Harre, R. (2000). Narrative: The Problems and Promises of One Alternative Paradigm. *Voprosy filosofii*, 3, 29–42. (In Russian)
- Chuganskaia, A. A. (2019). Scenario approach in assisting the buyer's economic behavior. In *Lomonosovsky readings — 2019. Economic Sciences Section. Economic relations in the context of digital transformation: a collection of theses of speeches* (pp. 439–442). Moscow: Moscow State University Publ. (In Russian)
- Dem'iankov, V. Z. (1994). Cognitive linguistics as a form of interpretive approach. *Voprosy iazykoznanii*, 4, 17–19. (In Russian)
- Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. 2nd ed. Berlin: Springer.
- Kusner, M., Sun, Y., Kolkin, N. I., Weinberger, K. Q. (2015). *From embeddings to document distances*. The 32nd International Conference on Machine Learning. Lille, France (pp. 957–966). Lille: ICML.
- Kutuzov, A., Kuzmenko, E. (2017). WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In D. Ignatov et al. (Eds), *Analysis of Images, Social Networks and Texts*. AIST 2016. Communications in Computer and Information Science, vol. 661 (pp. 155–161). Springer: Cham. DOI 10.1007/978-3-319-52920-2-15. Retrieved from <https://rusvectors.org/ru/>.
- Kuznetsova, Iu. M. (2018). Scenario approach to text analysis. *Works of ISA RAS*, 1 (68), 31–41. (In Russian)
- Kuznetsova, Iu. M. (2019). Social scenarios of behavior as a subject of online discussion. In *Category of "social" in modern pedagogy and psychology*. Materials of the 7th All-Russian scientific and practical conference with remote and international participation (pp. 177–183). Ufianovsk: Zebra Publ. (In Russian)
- Kuznetsova, Iu. M., Osipov, G. S., Smirnov, I. V., Chudova, N. V. (2019a). Text of network discussion as a source of script information. *Rechevye tekhnologii*, 1, 30–44. (In Russian)
- Kuznetsova, Iu. M., Suvorova, M. I., Chudova, N. V. (2019b). Scenario as a form of representation of an event in a landmark picture of the world. *Works of ISA RAS*, 1, 70–82. (In Russian)
- Kuznetsova, Iu. M., Chudova, N. V. (2008). *Psychology of Internet residents*. Moscow: LKI Publ. (In Russian)
- Leont'ev, A. N. (1974). *Activity. Consciousness. Personality*. Moscow: Politizdat Publ. (In Russian)
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013). Distributed Representations of Word and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119). Retrieved from <https://arxiv.org/abs/1310.4546>.
- Minskii, M. (1979). *Frames for presenting knowledge*. Moscow: Energiya Publ. (In Russian)
- Olker, Kh. R. (1987). Fairy tales, tragedies and the method of presenting the history of the world. In *Language and modeling of social interaction* (pp. 408–440). Moscow: Progress Publ. (In Russian)
- Osipov, G. S., Panov, A. I., Chudova, N. V., Kuznetsova, Iu. M. (2017). *Sign-based picture of the world of the subject behavior*. Moscow: Fizmatlit Publ. (In Russian)
- Osipov, G. S., Smirnov, I. V., Tikhomirov I. A. (2008). Relational-situational method of search and analysis of texts and its application. *Artificial intelligence and decision making*, 2, 3–10. (In Russian)
- Polatovskaia, O. S. (2013). Frame script as a type of concept. *Vestnik Irkutskogo gosudarstvennogo lingvisticheskogo universiteta*, 4 (25), 161–166. (In Russian)
- Propp, V. Ia. (1998). *Morphology of the (magic) fairy tale: historical roots of the magic fairy tale*. Moscow: Labyrinth Publ. (In Russian)
- Shliakhov, V. I. (2007). Scenarios of Russian speech interaction. *Russkii iazyk za rubezhom*, 2, 26–34. (In Russian)
- Smirnov, I. V., Shelmanov, A. O., Isakov, V. A., Stankevich, M. A. (2018). Open extraction of information from texts. Part I. Setting the task and reviewing the methods. *Artificial intelligence and decision making*, 2, 47–61. DOI: 10.14357/20718594180204. (In Russian)
- Vezhbitskaia, A. (1999). *Semantic universalia and description of languages*. Moscow: Languages of Russian culture Publ. (In Russian)
- Volosukhina, N. V. (2010). To the question of the interpretation of the concepts of "concept" and "frame" of modern linguistics. In *Materials of scientific and methodological readings of PGLU* (pp. 41–46). Piatigorsk: Pyatigorsky State Linguistic University Publ. (In Russian)

Received: October 1, 2019

Accepted: November 10, 2019

Authors' information:

Valerii A. Mishlanov — Dr. Sci. in Philology, Professor; mishl@psu.ru
 Anfisa A. Chuganskaya — PhD in Psychology; anfisa.makh@gmail.com
 Ivan V. Smirnov — PhD in Phys.-Math. Sci.; ivs@isa.ru
 Margarita I. Suvorova — Postgraduate Student; suvorova@isa.ru
 Ilya A. Kuruzov — Postgraduate Student; kuruzov2014@mail.ru