# Извлечение сценарной информации из текстов. Часть 1. Постановка задачи и обзор методов *

М. И. Суворова[I], М. В. Кобозева[I], Е. Г. Соколова[II], С. Ю. Толдова[II]

[I] Федеральный исследовательский центр «Информатика и управление» РАН, г.Москва, Россия
[II] Национальный исследовательский университет «Высшая школа экономики», г.Москва, Россия

**Аннотация.** В статье обсуждается важность автоматического сценарного анализа для понимания текстов на естественном языке. Дан широкий обзор методов и подходов к описанию и извлечению сценариев. Рассмотрены теоретические подходы к формализации сценариев. Приведен список задач, для решения которых используется информация о сценарной структуре текста. Представлены популярные подходы к автоматическому извлечению сценариев из текстов и методы оценки их качества. Показан список открытых корпусов, как созданных специально для задачи извлечения сценариев, так и универсальных, которые могут успешно применяться для данной задачи.

**Ключевые слова:** сценарный анализ, сюжетные грамматики, фрейм, скрипт, методы автоматической обработки естественного языка.

**DOI 10.14357/20718594200102**

## Введение

Понимание естественного языка до сих пор является одной из ключевых проблем искусственного интеллекта, эффективное решение которой невозможно без использования знаний о языке, его функционировании и строении, в особенности знаний о структуре текста и его составляющих. Поэтому при обработке текста широко применяется глубокий лингвистический анализ. Дискурсивный анализ – это одна из задач лингвистического анализа, объектом которой является целый текст, а не отдельные предложения (как в случае с синтаксисом и семантикой). Для его применения необходимо хотя бы несколько простых предложений. К дискурсивному анализу относится как установление семантических отношений между всеми клаузами текста (например, Теория риторических структур Манна и Томпсон [1]), так и извлечение из текста цепочек событий с их действующими лицами, расположенных в хронологическом порядке (например, narrative schemas Журавского и Чемберса [2]). Настоящее исследование является частью большого проекта, в котором для анализа текстов применяются оба подхода. Поэтому в данной работе будем использовать термин дискурсивный анализ для обозначения подхода с установлением семантических отношений между клаузами, а термин сценарный анализ – для извлечения цепочек событий.

Под сценарием будем понимать нечто близкое к понятию narrative schemas [2], а именно некоторую устоявшуюся (в социуме) последовательность шагов для достижения желаемой цели. При этом у каждого шага есть свои предусловия (несоответствие которым не позволит приступить к выполнению шага), действующие

*Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №17-29-07033.

✉ *Суворова Маргарита Игоревна. E-mail: suvorova@isa.ru*

лица (которым предписывается выполнить данное действие), а также операнды (объекты, с которыми оно выполняется). Сценарий содержит максимально полную информацию обо всех возможных путях развития описываемой ситуации (с точками выбора и ветвлениями). Поэтому он должен формироваться на основе целого корпуса, где каждый отдельный текст содержит некоторый сюжет, описывающий личный опыт автора по прохождению сценария или его частный случай. К примеру, сценарий «покупка автомобиля» состоит из трех основных шагов (выбор автомобиля, его осмотр и оформление покупки), каждый из которых представляет собой отдельный сценарий, со своими шагами и ветвлениями. У каждого шага есть своя цель (стать обладателем нового автомобиля, определиться с функционалом) и предусловия (не определившись с моделью и параметрами автомобиля нельзя переходить к его поиску и осмотру), действующие лица (покупатель, продавец, сотрудник ГИБДД) и операнды (автомобиль, договор купли-продажи).

Извлечение сценариев – это важный шаг на пути к истинному пониманию языка. Современные методы анализа и генерации текста, используемые, к примеру, в чат-ботах, системах поиска и сравнения текстов, достаточно плохо учитывают глобальную структуру текста. Один из распространенных способов заложить в модель информацию о глобальном содержании документа – описание тематики документа с помощью метода «мешок слов» или тематическое моделирование. При таком подходе теряется информация о структуре изложения, о взаимосвязях, упоминаемых в тексте концептов друг с другом. Качественные методы извлечения сюжетно-композиционного строения позволят существенно повысить качество решения многих других задач, включая синтез текста и рассуждения по тексту.

С точки зрения разработки прикладных систем зачастую хорошо работают комбинированные методы, когда низкоуровневые задачи решаются машинным обучением (извлечение признаков, извлечение фактов), а высокоуровневые (прикладные), задачи решаются системами правил. Сценарный анализ – это еще один этап анализа структуры текста, к результатам которого можно применять правила или другие методы решения конечной задачи. В то время как методы, основанные на непрерывных дифференцируемых моделях (например, искусственных нейронных сетях) показывают впечатляющее качество решения некоторых задач, их сложно комбинировать с другими подходами (например, системами правил). К тому же, такие модели строят векторные представления, которые сложно интерпретировать человеку («черный ящик»). Поэтому часто процесс анализа текста разбивают на несколько этапов (синтаксический и семантический разбор, извлечение сущностей и фактов, связывание сущностей и т.п.), а потом добавляют правила, чтобы на основе всего анализа решать конечную задачу – оценивать качество документа, близость текстов и т.д.

В статье приведен аналитический обзор задач, которые могут быть решены с помощью данных о структуре изложения, а также методов извлечения сценариев из текстов и описание существующих текстовых коллекций. Статья структурирована следующим образом: в первом разделе представлены теоретические подходы к описанию и формализации сценариев, во втором – задачи, для решения которых используется сценарный анализ. В третьем разделе речь идет о методах и подходах к автоматическому извлечению и построению сценариев, в четвертом – о методах оценки автоматических систем извлечения сценариев. В пятом разделе приводится список корпусов, которые можно использовать для разработки и тестирования систем.

## 1. Теоретические подходы к описанию и формализации сценариев

**Сюжетные грамматики**. Формализация структуры текста начиналась в виде анализа организации художественных текстов. «Фундаментальные принципы исследования сюжета художественного произведения были предложены В. Я. Проппом [3]. Его представления о сюжете волшебной сказки как комбинации функций персонажей легло в основу разнообразных логических моделей сюжета. Сюжетные тексты представляют последовательность действий с помощью отношения импликации между категориями сюжета, определяемыми шаблонами поведения персонажей» [4]. Изначально метод В. Я. Проппа был ориентирован на сказки, но он применим и к некоторым другим жанрам, например, басням. Ю.С. Мартемьянов предпринял попытку описать структуру басен в

виде глубинной структуры текста в терминах логического исчисления [5].

В литературе сюжет представляет собой законченную последовательность событий или умозаключений, объединенную внутренней логикой его развития. Первые теоретические работы по моделированию сюжета художественных произведений были направлены на обнаружение именно этой внутренней логики развития сюжета, которая непосредственно связана с жанром или типом текста. Как жанры и типы текстов, так и сюжетные модели этих текстов очень разнообразны. В то время как В.Я. Пропс использовал функции персонажей для моделирования сюжетов сказок, ученица Р. Шенка Ленерт построила модель сюжета новеллы О'Генри «Дары волхвов» из психологических состояний персонажей [6, с.29-30]. Исследовались также нехудожественные (специальные) тексты**,** например, инструкции, доказательства теорем и др. [6]. В частности, текст инструкции всегда представляет собой чередование процедур Цель и Метод, представляющих действия, осуществляемые адресатом текста инструкции, например: *Чтобы согреть воду в электрическом чайнике* (Цель адресата), *надо ее налить в чайник* (Метод – действие адресата) *и включить чайник* (Цель), *для этого надо нажать кнопку «on»* (Метод). Похожую структуру имеют и кулинарные рецепты.

В современной лингвистике понятие сюжета связывается в основном с самими художественными текстами, а не с жанрами. Согласно [4], современные исследования сюжета художественных произведений связаны с реконструкцией связей между персонажами с помощью подходов, основанных на правилах [7] и нейронных сетях [8].

**Скрипты и фреймы**. Более универсальное когнитивное направление связано с исследованиями по Искусственному интеллекту (ИИ), который занимается разработкой когнитивных процессуальных теорий, а также экспериментами по компьютерному воплощению этих теорий. Делая упор на процессуальное описание, ИИ занимается моделированием понимания. Понимание достигается объединенным применением семантических и прагматических знаний. Значение слова или высказывания представлено как составная часть памяти, точно так же, как и другие знания. Важнейшим понятием здесь является «*память*» [9]. Семантические теории должны использовать такие структуры для представления обыденных (common-sense) знаний, как *фреймы* [10,11] *и сценарии (*или *скрипты* – scripts) [12]. В работе Филлмора [13] дается особенно четкий пример такого подхода: «В английском языке есть семантическая область, связанная с тем, что можно назвать «торговой сделкой». Фрейм ее имеет форму сценария (scenario), содержащего роли, которые можно обозначить как 'покупатель', 'продавец', 'товар' и 'деньги'; он содержит такие «кадры», в которых показано, как покупатель отдает деньги и берет товар, а продавец отдает товар и берет деньги. Весь сценарий (scenario) торговой сделки активизируется в сознании всякий раз, когда человек встречает и понимает любое из следующих слов: buy 'покупать', sell 'продавать', pay 'платить', cost 'стоить', spend 'тратить', charge 'назначить цену' и т.д., хотя каждое из них выделяет или выводит на первый план, только небольшой участок фрейма. Каждое из этих слов несет в себе одновременно и фон, и изображение, и всю обстановку, и ту часть этой обстановки, на которую указывает данное слово [9].

Основные понятия, возникшие в этом направлении и получившие широкое распространение, – это *фрейм* и *скрипт* (часто переводится как «сценарий»), причем первый использовался в основном для представления статической информации, например, фрейм «комната», а сценарий или скрипт – для процессуальной, представляющей обычную последовательность событий, например, «посещение ресторана». Скрипт – это некоторый набор ожиданий о том, что в воспринимаемой ситуации должно произойти дальше [14], устоявшаяся в культуре последовательность шагов, «знание того, как поступать и как другие поступят в конкретных стереотипных ситуациях» [15]. Практически эти понятия стали использоваться для генерации и анализа текстов вне направления ИИ, в частности, в генерации текстов, для моделирования таких стереотипных текстов, как, например, анкета, биография, сводка погоды. К этому же типу можно отнести и более специальные сценарии, например, когнитивные схемы «медицины страдающих органов», которые позволяют врачу предложить пациенту для обсуждения сценарий его лечения [16].

**Нарративные схемы**. Журавский и Чемберс в работе [17] предлагают рассматривать сценарий (narrative schemas) как последовательность взаи-

мосвязанных событий (например, арестовать, осудить), аргументы которых заполнены семантическими ролями участников (полиция, судья, обвиняемый). Данный подход основывается на понятии скрипта. Формально, сценарий представляется в виде графа, в вершинах которого стоят действия вместе с ожидаемыми аргументами, а рёбра соответствуют отношению следования. При этом граф должен содержать все возможные действия, которые потенциально могут возникнуть в рассматриваемой жизненной ситуации. Очевидно, такой граф невозможно построить по одному тексту, поэтому авторы предлагают анализировать коллекцию текстов, посвящённых одной ситуации.

## 2. Задачи, для которых используется информация о структуре изложения

С 70-х – 80-х годов прошлого века сценарии стали центральной темой в исследовании понимания естественного языка, в частности решения задач реферирования, вопросно-ответного поиска и разрешения кореференции [18]. В настоящее время сценарный анализ применяется для решения различных задач. Перечислим их основные типы.

**Извлечение информации из текста**. Отметим, что в данном случае речь идет об извлечении информации именно на уровне текста, в то время как большинство исследований в этой области ограничивается рамками одного предложения – могут извлекаться как отдельные именованные сущности, так и семантические отношения на уровне синтаксиса [19-20].

Если в большинстве работ по извлечению информации выделяются отдельные события или аргументы, то нарративные схемы помогают извлекать события, так или иначе соотносящиеся между собой [2] (например, такие два события как редактирование текста и издание книги). В работе [21] предлагается алгоритм, который извлекает информацию без предварительного описания шаблонов. Он автоматически распознает структуру текста из необработанных данных и предлагает цепочки, представляющие собой набор связанных событий (например, бомбардировка включает в себя взрыв и разрушение), сопряженных с семантическими ролями. Для *извлечения аргументации и логических выводов* из текста могут использоваться нарративные схемы. Они не являются достаточным решением, но их можно рассматривать как шаг к изучению причинно-следственных отношений [2].

Подходы к *автоматическому реферированию* на основе знаний о структуре повествования показывают лучшие результаты, чем подходы на основе токенов или даже более сложные подходы. Так, в работе [2] нарративные схемы помогают устанавливать связи между предикатами в документе. В подобных подходах моделируется отдельное событие, упомянутое в тексте, и определяется насколько оно важно для заданной темы [22, 23].

Еще в 1982 году исследователями из Йельского университета была разработана система FRUMP [24], действующая на основе сценариев разных событий, которые обычно описываются в новостных статьях (политика и дипломатические отношения, демонстрации, природные и техногенные катастрофы, террористические акты и т.д.). Для системы заранее составлялись сценарии, включающие только самые ключевые этапы событий (например, задержание преступников: полиция выехала на место преступления, перестрелка между полицией и преступниками, полиция задержала преступников, преступники предстали перед судом, преступников заключили под стражу). Затем система анализировала текст, выбирала конкретный сценарий, к которому он относится, подставляя в него действующие лица, локации и время событий из текста, и проверяя, какие именно шаги из данного сценария встретились в тексте.

В последние годы появляются попытки объединить современные нейросетевые методы анализа текстов (BERT, transfer learning) и классический структурный подход к анализу дискурса RST для решения задачи реферирования текстов [25].

**Разрешение кореференции.** В работе [26] описывается система разрешения кореференции (BABAR), основанная на информации о структуре повествования – «фреймы ситуаций» (Caseframe Network). Фреймы ситуаций – это пара предикатов, которые имеют один и тот же аргумент (X заявил, X добавил). Учет связей между такими аргументами может повышать качество разрешения кореференции. В работе [27], наоборот, нарративные схемы извлекаются с опорой на информацию о кореференции (идентифицируются предикаты, чьи аргументы находятся в отношениях кореференции).

В работе [28] показано, что структура изложения может также использоваться для *различения типов новостных документов***:** нарративно гомогенные типы текстов более жесткие по своей структуре, чем нарративно неоднородные. В них используется ограниченный набор схем и структура их повествования очень предсказуема. К гомогенному типу в данном исследовании отнесены категории «Свадьбы» и «Некрологи». Напротив, неоднородные типы текстов имеют самые разные нарративные структуры, что объясняется разнообразием их содержания. Так, события в категории «Вооружение и оборона США» могут сильно различаться – от взрывов на дорогах до перерасхода бюджета.

Кроме того, сценарный анализ может использоваться для более специфичных задач. Так, в работе [29] при помощи «нитей сюжета» решается задача отслеживания *нагнетания тревожности в произведении***.** Модель может использоваться для прогнозирования реакции читателя еще на этапе генерации нарратива.

## 3. Методы и подходы к автоматическому извлечению и построению сценариев

Существуют подходы, которые с точки зрения методов обработки естественного языка (NLP) можно назвать ***классическими***. К таким методам относится работа [18]. С помощью библиотеки OpenNLP авторы извлекли из текстов цепочки вида Глагол-Субъект и Глагол-Объект, а также кореферентные связи между ними. Затем из всех полученных пар отбирались статистически устойчивые цепочки (т.е. действие А и действие Б часто встречаются вместе с одним действующим лицом). Для этого использовалась метрика связанности событий PMI (точечная взаимная информация). На следующем этапе все действия выстраивались в хронологическом порядке с помощью классификатора, обученного на корпусе TimeBank. Для объединения цепочек событий, извлеченных из разных текстов, применялись агломеративная иерархическая кластеризация.

Авторы работы [30] взяли за основу подход, описанный выше, и улучшили его с помощью открытого извлечения информации (OLLIE Open IE). Из текстов извлекались триплеты вида Субъект-Глагол-Объект, которые затем обобщались (к примеру, на место субъекта выбиралось наиболее общее понятие вместо имени собственного). Далее с помощью языковой модели оценивалась вероятность совместного возникновения триплетов, и строился взвешенный граф, из которого удалялись слабые связи.

К другому классу методов относятся ***нейросетевые методы***. В частности, рекуррентные нейросети (LSTM), которые предсказывают следующий элемент последовательности по текущему. Так, в работе [31] каждая клауза текста преобразуется в кортеж из пяти элементов (глагол, субъект, объект и еще максимум 2 аргумента) и затем совместное распределение кортежей в последовательности моделируется с помощью LSTM. В работе не упоминаются языковые модели, хотя по сути это тесно связанная задача. Авторы рассматриваемой статьи получили не очень хорошее качество, что вполне ожидаемо в виду небольшого размера корпуса и отсутствия какого-либо переноса навыков (transfer learning) и использования предобученных языковых моделей. Есть основания полагать, что совмещение данного подхода с применением современных нейронных сетей позволит существенно улучшить показатели качества.

В литературе также встречаются и более ***экзотические подходы*** к извлечению сценариев. Например, в работе [32] анализировались тексты на японском языке, поэтому и методы применялись другие (отличные от исследований для английского языка). В частности, авторы извлекали пары Предикат-Аргумент и объединяли их в устойчивые сочетания с помощью алгоритма Априори [33].

Другим примером необычного подхода можно назвать работу [34]. Авторы использовали корпус текстов, собранных через Amazon Mechanical Turk. Волонтерам ставилась задача описать сценарии некоторых бытовых ситуаций в виде нумерованного списка шагов (как у Шенка: зашел в ресторан, сделал заказ, оплатил счет и т.д.). Затем авторы сопоставляли списки разных участников по одним и тем же тематикам. Для этого применялся алгоритм из области биоинформатики (Multiple Sequence Alignment Algorithm), который сопоставляет элементы нескольких последовательностей, не переставляя их местами. Семантическая близость элементов списков измерялась с помощью онтологии WordNet.

## 4. Методы для оценивания качества моделей

Разработка эффективных методов невозможна без соответствующих метрик качества, которые должны быть одновременно прозрачными и надежными.

Рассмотрим два самых популярных подхода:

1. *Narrative Cloze* – способ, при котором из цепочки событий удаляется случайный элемент, а компьютер должен заполнить пробел (метод предложен в [35] для оценки того, как хорошо человек понимает текст). Применялся в работе [17].

2. *Story Cloze Test*, который заключается в том, чтобы для короткой истории в 4 предложения выбрать наиболее подходящее завершение из двух предложенных вариантов [36]. Эта задача тесно связана с проблемой определения противоречивости двух фрагментов текста (textual entailment).

Пример теста:

*Контекст*. Сара мечтала посетить Европу в течение многих лет. Она наконец-то накопила достаточно средств для поездки. Она приземлилась в Испании и совершила путешествие на восток вглубь континента. Ей не понравилось, как все отличалось (от привычной для нее обстановки).

*Правильное окончание текста*: Сара поняла, что ей больше нравится свой дом, чем Европа.

*Неправильное окончание текста*: Сара решила переехать в Европу.

Авторы [37] метрики предложили несколько базовых методов (baseline) как ориентиров для дальнейшего сравнения:

- наличие более частотных для языка слов (для отсечения не очень вероятных в реальном мире вариантов);
- перекрытие по n-граммам с началом истории;
- близость по word2vec [37];
- соответствие по тональности первому/последнему предложению истории;
- Skip-thoughts [38] – метод подобный word2vec, но использующий рекуррентные нейросети;
- метод Журавского через PMI;
- нейросетевой DSSM метод [39], который применяется для поиска похожих объектов.

Одна из очевидных проблем данной метрики заключается в том, что выбор делается из двух вариантов. Таким образом, выбирая каждый раз, к примеру, только первый вариант, можно решить задачу примерно с 50%-ным качеством. Значения метрики, приведенные в статье, попадают в диапазон от 0,47 до 0,59.

В работе [34] для оценки качества предлагалось использовать детектирование парафраза, т.е. способность системы распознать описание того же самого события (для которого уже построен граф), сформулированного другими словами.

Часто оценка качества работы автоматическими метриками дополняется опросами с участием людей (в первую очередь, через crowdsourcing-платформы – Amazon Mechanical Turk и Яндекс.Толока). При этом автоматические метрики дешевле и могут быть использованы для предварительной настройки параметров модели, а финальная оценка может быть проведена с привлечением волонтеров [30, 32].

## 5. Размеченные корпуса

На современном этапе развития области одна из самых фундаментальных проблем – необходимость разметки больших корпусов для настройки моделей, основанных на машинном обучении с учителем. Один из наиболее популярных подходов для борьбы с нехваткой размеченных данных – перенос навыков (transfer learning), когда искусственная нейросеть «предобучается» на задаче, где не требуется ручная разметка (например, моделирование языка, language modeling [40]). В результате такая нейросеть может строить информативные векторные представления текстов и слов и относительно легко адаптируется для решения более специфической задачи в режиме обучения с учителем, при этом достаточно относительно небольшого размеченного корпуса. Существуют также подходы, основанные на активном обучении и few-shot learning [41- 43].

В открытом доступе также имеются готовые достаточно объемные корпуса, которые могут помочь в работе со сценариями, но все они англоязычные. К ресурсам, созданным специально для извлечения сценарной информации, относятся следующие три корпуса.

1. Авторы работы [37] собрали большой ***корпус коротких историй*** (из пяти предложений) на разные тематики с помощью платформы Amazon Mechanical Turk, а затем проверили каждую историю вручную. В итоге у них оста-

лось 49 255 уникальных историй. Данный корпус может служить ресурсом для изучения структуры повествования, а также для обучения моделей генерации текстов. Корпус выложен в открытый доступ[1].

**2**. ***База уже извлеченных цепочек взаимосвязанных событий*** (с действующими лицами), расположенных в хронологическом порядке, доступна вместе с другими интересными корпусами на страничке одного из авторов описанной выше работы - Найта Чемберса (Nate Chambers)[2]. Корпус разбит на несколько кластеров в зависимости от длины цепочек.

3. Волонтерам с Amazon Mechanical Turk было предложено ***описать 22 сценария*** (свадьба, создание сайта, приготовление яичницы, поход в ресторан быстрого питания…) в виде списка действий, включающего от 5 до 16 пунктов расположенных в хронологическом порядке. В итоге авторы [34] получили почти 500 сценариев от 25 волонтеров. Страничка проекта давно не поддерживается, но еще работает[3].

К более универсальным корпусам, которые также используются в некоторых работах по анализу сценарной информации, относятся следующие ресурсы.

• В работе [28] использовался ***The New York Times Annotated Corpus*** [44], содержащий более 1,8 млн статей, опубликованных в газете “The New York Times” с 01.01.1987 по 19.07.2007 гг. Каждая статья включает метаданные, вручную составленную аннотацию, тематические метки, размеченные именованные сущности (люди, организации, географические наименования). Корпус доступен в сети[4].

• Другим корпусом, который также доступен на Linguistic Data Consortium, является ***Annotated English Gigaword***[5]. В корпусе содержится почти 10 млн документов, включающих разбивку на слова и токены, автоматически сгенерированные аннотации, синтаксическую разметку, выделенные именованные сущности, кореферентные цепочки. Кроме того, для корпуса Gigaword есть также дополнительная разметка, подготовленная для соревнования Narrative Cloze Evaluation[6].

• Для организации событий в хронологически упорядоченные цепочки может быть полезен корпус ***The Timebank Corpus*** [45], в котором размечены все события, а также установлены бинарные отношения между ними, указывающие на временной порядок (до, после, включает, одновременно и др.). Корпус доступен в Интернете[7].

## Заключение

В статье рассмотрены теоретические подходы к описанию сценариев, а также современные методы автоматического извлечения сценарной структуры текстов, позволяющие решать многие задачи, связанные с пониманием естественного языка. Представлены также корпуса для обучения и тестирования систем и подходы к оценке качества их работы.

Во второй части этой работы будут проведены экспериментальные исследования методов автоматического извлечения сценариев на коллекции русскоязычных текстов.

Несмотря на то, что современные методы показывают хорошие результаты для решения многих задач, не стоит забывать и про классические статистические методы анализа текстов. Планируется применить метод Журавского (D. Jurafsky) для текстов на русском языке: выделять различные связи, а затем отделять устойчивые связи с помощью современных методов, в том числе нейросетевых.

## Литература


1. Mann W. C., Thompson S. A. Rhetorical structure theory: Toward a functional theory of text organization //Text-interdisciplinary Journal for the Study of Discourse. 1988. Т. 8. №. 3. P.243-281.
2. Chambers N., Jurafsky D. A Database of Narrative Schemas //LREC. 2010.
3. Пропп В. Я. Морфология (волшебной) сказки: исторические корни волшебной сказки. М.: Лабиринт. 1998.
4. Митрофанова О.А. Исследование структурной организации художественного произведения с помощью тематического моделирования: опыт работы с текстом романа «Мастер и Маргарита» М.А. Булгакова // Труды международной конференции «Корпусная лингвистика-2019» СПГУ. СПб. 2019. С.387-394.
5. Мартемьянов Ю. С. Логика ситуаций //Строение текста. Терминологичность слов. М.: ЯСК. 2004.
6. Баранов А.Н. Введение в прикладную лингвистику. М.: Эдиториал УРСС. 2001.


[1] http://cs.rochester.edu/nlp/rocstories/
[2] https://www.usna.edu/Users/cs/nchamber/
[3] http://www.coli.uni-saarland.de/projects/smile/
[4] https://catalog.ldc.upenn.edu/LDC2008T19
[5] https://catalog.ldc.upenn.edu/LDC2012T21
[6] https://www.usna.edu/Users/cs/nchamber/data/chains/
[7] http://www.timeml.org/timebank/timebank.html

7. Bodrova A. A., Bocharov V. V. Relationship Extraction from Literary Fiction //Dialogue: International Conference on Computational Linguistics. 2014.
8. Iyyer M. et al. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships //Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016. P.1534-1544.
9. Шенк Р., Бирнбаум Л. Мей Дж. К интеграции семантики и прагматики //Новое в зарубежной лингвистике. 1989. №. 24. С. 32-47.
10. Minsky M. L. Frame-system theory //Thinking. 1977.
11. Charniak E. On the use of framed knowledge in language comprehension //Artificial Intelligence. 1978. Т. 11. №. 3. P. 225-265.
12. Schank R. C., Abelson R. P. Scripts //Plans, Goals and Understanding. 1977.
13. Fillmore C. J. et al. Frame semantics and the nature of language //Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech. 1976. Т. 280. №. 1. P.20-32.
14. Schank R. C., Abelson R. P. Scripts, plans, and knowledge //IJCAI. 1975. P. 151-157.
15. Дарбанов Б. Е. Теория схемы, фрейм, скрипт, сценарий как модели понимания текста //Актуальные проблемы гуманитарных и естественных наук. 2017. №. 6-2. С. 75-78.
16. Тхостов А.Ш. Нелюбина А.С. Обыденные представления о болезни в структуре идентификации пациента и врача как предиктор выбора пациентом способа лечения (на модели сердечно-сосудистых заболеваний // Общество ремиссии: на пути к нарративной медицине: сб. науч. тр. – Самара: Самарский университет. 2012. С.12-33.
17. Chambers N., Jurafsky D. Unsupervised learning of narrative schemas and their participants //Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. – Association for Computational Linguistics. 2009. P.602-610.
18. Chambers N., Jurafsky D. Unsupervised learning of narrative event chains //Proceedings of ACL-08: HLT. 2008. P.789-797.
19. Козеренко Е. Б., Кузнецов К. И., Романов Д. А. Семантическая обработка неструктурированных текстовых данных на основе лингвистического процессора PullEnti //Информатика и ее применения. 2018. Т. 12. №. 3. P. 91-98.
20. Шелманов А. О. и др. Открытое извлечение информации из текстов. Часть I. Постановка задачи и обзор методов //Искусственный интеллект и принятие решений. 2018. №. 2. С. 47-61.
21. Chambers N., Jurafsky D. Template-based information extraction without the templates //Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics. 2011. P.976-986.
22. Azzam S., Humphreys K., Gaizauskas R. Using coreference chains for text summarization //Proceedings of the Workshop on Coreference and its Applications. – Association for Computational Linguistics. 1999. P.77-84.
23. Filatova E., Hatzivassiloglou V. Event-based extractive summarization //Text Summarization Branches Out. 2004. P.104-111.
24. DeJong G. An overview of the FRUMP system //Strategies for natural language processing. 1982. Т. 113. P.149-176.
25. Xu J. Gan Z., Cheng Y., Liu J. Discourse-Aware Neural Extractive Model for Text Summarization //arXiv preprint arXiv. 1910.14142. 2019.
26. Bean D., Riloff E. Unsupervised learning of contextual role knowledge for coreference resolution //Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. 2004. P.97-304.
27. Irwin J., Komachi M., Matsumoto Y. Narrative schema as world knowledge for coreference resolution //Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. – Association for Computational Linguistics. 2011. P.86-92.
28. Simonson D., Davis A. NASTEA: Investigating narrative schemas through annotated entities //Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016). 2016. P.57-66.
29. Doust R., Piwek P. A model of suspense for narrative generation //Proceedings of the 10th International Conference on Natural Language Generation. 2017. P.178-187.
30. Balasubramanian N. et al. Generating coherent event schemas at scale //Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013. P.1721-1731.
31. Pichotta K., Mooney R. J. Learning statistical scripts with LSTM recurrent neural networks //Thirtieth AAAI Conference on Artificial Intelligence. 2016.
32. Shibata T., Kohama S., Kurohashi S. A Large Scale Database of Strongly-related Events in Japanese //LREC. 2014. P.3283-3288.
33. Borgelt C., Kruse R. Induction of association rules: Apriori implementation //Compstat. – Physica, Heidelberg. 2002. P.395-400.
34. Regneri M., Koller A., Pinkal M. Learning script knowledge with web experiments //Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 2010. P.979-988.
35. Taylor W. L. “Cloze procedure”: A new tool for measuring readability //Journalism Bulletin. 1953. Т. 30. №. 4. P.415-433.
36. Mostafazadeh N. et al. A corpus and evaluation framework for deeper understanding of commonsense stories //arXiv preprint arXiv:1604.01696. 2016.
37. Mikolov T. et al. Distributed representations of words and phrases and their compositionality //Advances in neural information processing systems. 2013. P. 3111-3119.
38. Kiros R. et al. Skip-thought vectors //Advances in neural information processing systems. 2015. P.3294-3302.
39. Huang P. S. et al. Learning deep structured semantic models for web search using clickthrough data //Proceedings of the 22nd ACM international conference on Information & Knowledge Management. – ACM. 2013. P.2333-2338.
40. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. 2018.
41. Settles B. Active learning //Synthesis Lectures on Artificial Intelligence and Machine Learning. 2012. Т. 6. №. 1. P.1-114.
42. Suvorov R., Shelmanov A., Smirnov I. Active Learning with Adaptive Density Weighted Sampling for Information Extraction from Scientific Papers //Conference on

Artificial Intelligence and Natural Language. Springer, Cham 2017. P.77-90.
43. Snell J., Swersky K., Zemel R. Prototypical networks for few-shot learning //Advances in Neural Information Processing Systems. 2017. P.4077-4087.
44. Sandhaus, E. The New York Times Annotated Corpus. Linguistic Data Consortium. Philadelphia. 2008.
45. Pustejovsky J. et al. The timebank corpus //Corpus linguistics. 2003. T.2003. 40 p.

## Extraction of Script Knowledge from Texts. Part I. The Task and the Review of the State of the Art

M. I. Suvorova[I], M. V. Kobozeva[I], E. G. Sokolova[II], S. Y. Toldova[II]

[I] Federal Research Center "Informatics and Control" of Russian Academy of Sciences, Moscow, Russia
[II] National Research University Higher School of Economics, Moscow, Russia

**Abstract**: This paper discusses the importance of automatic extraction of script knowledge for natural language understanding. We discuss theoretical approaches to the description of text structure: story grammars, scripts, frames and narrative schemas. We provide a list of research fields where automatic script knowledge extraction can be applied to achieve better precision and recall (e.g. automatic summarization, information extraction, coreference resolution, etc.). The article also presents popular approaches to the automatic extraction of script knowledge and methods for evaluation of such approaches. Besides, we present a list of datasets that can be used to train and test new models.
**Keywords:** script knowledge extraction, narrative schemas, scripts, frames, natural language processing.

**DOI** 10.14357/20718594200102

## References

1. Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. Text-interdisciplinary Journal for the Study of Discourse, 8(3), 243-281.
2. Chambers N., Jurafsky D. A Database of Narrative Schemas //LREC. 2010.
3. Propp V. Morphology of the Folktale. – University of Texas Press, 2010. – T. 9.
4. Mitrofanova O. 2019. Issledovaniye strukturnoy organizatsii khudozhestvennogo proizvedeniya s pomoshch'yu tematicheskogo modelirovaniya: opyt raboty s tekstom romana «Master i Margarita» M.A. Bulgakova [A study of the structural organization of fiction using topic modeling: the case study of the novel "The Master and Margarita" by Bulgakov] // Trudy mezhdunarodnoy konferentsii «Korpusnaya lingvistika-2019» [Proceedings of the international conference "Corpus Linguistics-2019"]. SPb. 387-394.
5. Martem'yanov Yu. 2004. Logika situatsiy [Logic of situations] // Stroyeniye teksta. Terminologichnost' slov. [Text structure. Terminology of words]. Moscow: YASK.
6. Baranov A.N. 2001. Vvedeniye v prikladnuyu lingvistiku [Introduction to Applied Linguistics]. Moscow, Editorial URSS.
7. Bodrova A. A., Bocharov V. V. 2014. Relationship Extraction from Literary Fiction //Dialogue: International Conference on Computational Linguistics.
8. Iyyer, M., Guha, A., Chaturvedi, S., Boyd-Graber, J., & Daumé III, H. (2016, June). Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 1534-1544.
9. Shenk, R., Birnbaum, L., and May, J. (1989). K integracii semantiki i pragmatiki [Integrating Semantics and Pragmatics], translated by G.Y. Levin, Novoe v zarubezhnoj lingvistike, 24, 32–47 (in Russian).
10. Minsky M. L. 1977. Frame-system theory //Thinking.
11. Charniak E. 1978. On the use of framed knowledge in language comprehension //Artificial Intelligence. (11: 3). 225-265.
12. Schank R. C., Abelson R. P. 1977. Scripts //Plans, Goals and Understanding.
13. Fillmore C. J. et al. Frame semantics and the nature of language //Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech. – 1976. – T. 280. – №. 1. – C. 20-32.
14. Schank R. C., Abelson R. P. Scripts, plans, and knowledge //IJCAI. – 1975. – C. 151-157.
15. Darbanov B. 2017. Teoriya skhemy, freym, skript, stsenariy kak modeli ponimaniya teksta [Scheme theory, frame, script, script as a model for understanding of the text] // Aktual'nyye problemy gumanitarnykh i yestestvennykh nauk [Actual problems of the humanities and natural sciences]. No. 6-2. 75-78.
16. Tkhostov, A., & Nelyubina, A. 2013. Illness Perceptions in Patients with Coronary Heart Disease and Their Doctors. Procedia-Social and Behavioral Sciences. 86: 574-577.
17. Chambers N., Jurafsky D. 2009. Unsupervised learning of narrative schemas and their participants //Proceedings of the Joint Conference of the 47th Annual Meeting of the

ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. – Association for Computational Linguistics. 602-610.
18. Chambers N., Jurafsky D. 2008. Unsupervised learning of narrative event chains //Proceedings of ACL-08: HLT. 789-797.
19. Kozerenko Ye. B., Kuznetsov K. I., Romanov D. A. 2018. Semanticheskaya obrabotka nestrukturirovannykh tekstovykh dannykh na osnove lingvisticheskogo protsessora PullEnti [Integrated Platform For Multilingual Text Knowledge Processing] // Informatika i yeyo primeneniya ["Informatics and Applications" scientific journal]. 12:3. 91-98.
20. A.O. Shelmanov, V.A. Isakov, M.A. Stankevich, I.V. Smirnov. 2018. Otkrytoye izvlecheniye informatsii iz tekstov Chast' I. Postanovka zadachi i obzor metodov [Open Information Extraction. Part I. The Task and the Review of the State of the Art ] // Iskusstvennyy intellekt i prinyatiye resheniy [Artificial Intelligence and Decision Making]. 2:47-61.
21. Chambers N., Jurafsky D. 2011. Template-based information extraction without the templates //Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics. 976-986.
22. Azzam S., Humphreys K., Gaizauskas R. 1999. Using coreference chains for text summarization //Proceedings of the Workshop on Coreference and its Applications. – Association for Computational Linguistics.77-84.
23. Filatova E., Hatzivassiloglou V. 2004. Event-based extractive summarization //Text Summarization Branches Out. 104-111.
24. DeJong G. 1982. An overview of the FRUMP system //Strategies for natural language processing. 113:149-176.
25. Xu J. Gan, Z., Cheng, Y., Liu, J. Discourse-Aware Neural Extractive Model for Text Summarization //arXiv preprint arXiv:1910.14142. 2019.
26. Bean D., Riloff E. 2004. Unsupervised learning of contextual role knowledge for coreference resolution //Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. 297-304.
27. Irwin J., Komachi M., Matsumoto Y. Narrative schema as world knowledge for coreference resolution //Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. – Association for Computational Linguistics, 2011. – C. 86-92.
28. Simonson D., Davis A. 2016. NASTEA: Investigating narrative schemas through annotated entities //Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016). 57-66.
29. Doust R., Piwek P. 2017. A model of suspense for narrative generation //Proceedings of the 10th International Conference on Natural Language Generation. 178-187.
30. Balasubramanian N. et al. 2013. Generating coherent event schemas at scale //Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 1721-1731.
31. Pichotta K., Mooney R. J. 2016. Learning statistical scripts with LSTM recurrent neural networks //Thirtieth AAAI Conference on Artificial Intelligence.
32. Shibata T., Kohama S., Kurohashi S. 2014. A Large Scale Database of Strongly-related Events in Japanese //LREC. 3283-3288.
33. Borgelt C., Kruse R. 2002. Induction of association rules: Apriori implementation //Compstat. – Physica, Heidelberg. 395-400.
34. Regneri M., Koller A., Pinkal M. 2010. Learning script knowledge with web experiments //Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. 979-988.
35. Taylor W. L. 1953. "Cloze procedure": A new tool for measuring readability //Journalism Bulletin. 30: 4. 415-433.
36. Mostafazadeh N. et al. A corpus and evaluation framework for deeper understanding of commonsense stories //arXiv preprint arXiv:1604.01696. 2016.
37. Mikolov T. et al. 2013. Distributed representations of words and phrases and their compositionality //Advances in neural information processing systems. 3111-3119.
38. Kiros R. et al. 2015. Skip-thought vectors //Advances in neural information processing systems. 3294-3302.
39. Huang P. S. et al. Learning deep structured semantic models for web search using clickthrough data //Proceedings of the 22nd ACM international conference on Information & Knowledge Management. ACM, 2013. 2333-2338.
40. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. 2018.
41. Settles, B. (2012). Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 6(1), 1-114.
42. Suvorov R., Shelmanov A., Smirnov I. 2017. Active Learning with Adaptive Density Weighted Sampling for Information Extraction from Scientific Papers //Conference on Artificial Intelligence and Natural Language. Springer, Cham. 77-90.
43. Snell J., Swersky K., Zemel R. 2017. Prototypical networks for few-shot learning //Advances in Neural Information Processing Systems. 4077-4087.
44. Sandhaus, E. 2008. The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia.
45. Pustejovsky J. et al. 2003. The timebank corpus //Corpus linguistics. 40p.