# Компьютерный анализ текстов

## Обучение анализатора для определения ролевых структур высказываний в текстах на русском языке на автоматически размеченном корпусе*

А.О. Шелманов, М.А. Каменская

**Аннотация.** В работе исследованы подходы к определению ролевых структур высказываний, использующие принципы машинного обучения с частичным привлечением учителя. Представлен способ повышения качества семантического анализа за счет обучения на корпусе, автоматически размеченном словарным (основанным на правилах) семантическим анализатором. Предложен подход к определению ролевых структур высказываний для «неизвестных» предикатных слов, которые отсутствуют в семантическом словаре словарного анализатора. В работе также представлен гибридный семантический анализатор, в котором используется две модели машинного обучения для «известных» и «неизвестных» предикатных слов, а также словарный семантический анализатор. Проведены экспериментальные исследования на вручную размеченном русскоязычном корпусе, которые показывают, что предложенные модификации повышают полноту и общее качество определения ролевых структур высказываний.

**Ключевые слова:** *определение ролевых структур высказываний, машинное обучение с частичным привлечением учителя, семантический анализ, векторные представления слов.*

### Введение

Определение ролевых структур высказываний (semantic role labeling) [1–3] является одной из разновидностей семантического анализа. Она также известна под названием распознавание семантических ролей [4]. В этой задаче семантика предложений текстов на естественном языке моделируется с помощью предикатных слов (ПС) (глаголов, отглагольных существительных, причастий и др.), обозначающих в тексте ситуации, их аргументов – синтаксических конструкций, обозначающих участников ситуаций, и значений аргументов – семантических ролей, которые играют участники в ситуации. Стоит также отметить, что предикатные слова с одинаковым значением могут группироваться во фреймы (словарные статьи), в рамках которых имеется единая ролевая структура. На рис. 1 представлен пример ролевой структуры предложения.

Такая модель текста позволяет абстрагироваться от морфологической и синтаксической структуры предложения и разным грамматическим конструкциям ставить в соответствие одинаковые семантические роли. Это, например, позволяет сравнивать предложения по смыслу, игнорируя различия их поверхностных структур. Ролевые структуры высказываний используются во многих прикладных задачах: в вопросно-ответном поиске [5–7], машинном переводе [8], оценке смысловой близости фрагментов текстов на естественном языке [9] и др.

Традиционный подход к решению задачи определения ролевых структур высказываний заключается в представлении ее в виде нескольких подзадач классификации, которые решаются методами машинного обучения с учителем (supervised machine learning). Этот подход впервые был предложен в работе [2] и развит многими другими исследователями. Например, в [10] усовершенствовано пространство признаков, в [11] рассмотрены различные методы классификации, в [12] иссле-

* Работа выполнена при поддержке РФФИ, проект №16-37-00425 «мол_а»

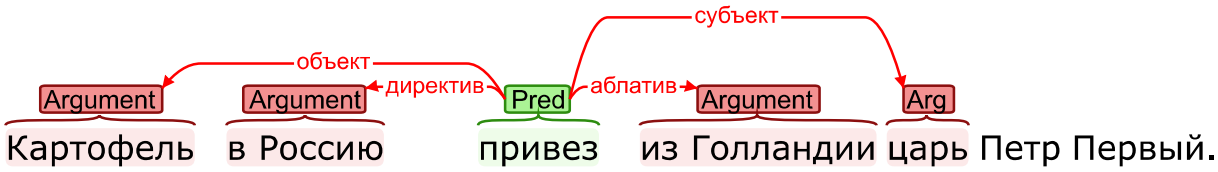


**Рис. 1.** Пример ролевой структуры предложения, полученной с помощью семантико-синтаксического анализатора [10]. Семантические роли: «объект» – то, на что направлено действие; «субъект» – инициатор действия; директив – «конечная точка действия»; «аблатив» – исходная точка действия»

дованы различные способы декомпозиции задачи определения ролевых структур высказываний на подзадачи, а в [13] предложены различные способы учета лингвистических ограничений на предикатно-аргументные структуры и взаимосвязей между аргументами.

Недостатком такого подхода является большая зависимость анализаторов от размера и качества семантически размеченных обучающих корпусов. Чтобы обеспечить приемлемое качество анализа, обучающие корпуса должны быть весьма объемными. Для английского языка существуют достаточно крупные ресурсы: PropBank [14] и FrameNet [15]. Однако создание подобных ресурсов очень трудоемкая и дорогостоящая работа. Поэтому для многих языков корпуса, пригодные для машинного обучения анализаторов, решающих задачу определения ролевых структур высказываний, отсутствуют. Стоит также отметить, что при анализе текстов из одной предметной области анализатором, обученным на корпусе из другой предметной области, качество анализа заметно снижается [16]. Таким образом, чтобы эффективно применять методы машинного обучения с учителем при решении задачи определения ролевых структур высказываний, необходимо не только создавать корпуса для интересующего языка, но и адаптировать их для каждой интересующей предметной области, что весьма трудоемко и неэффективно.

В настоящем исследовании была поставлена цель снизить объем ручной разметки корпусов текстов на естественном языке при создании семантических анализаторов для определения ролевых структур высказываний, основанных на машинном обучении, разработать способ адаптации семантических анализаторов к новым предметным областям и повысить качество семантического анализа текстов на естественном языке. Для этого были исследованы методы определения ролевых структур высказываний, основанные на машинном обучении с частичным привлечением учителя (semi-supervised machine learning). Это направление объединяет в себе подходы, ориентированные либо на формирования размеченного корпуса, пригодного для машинного обучения, либо непосредственно на создание обученного анализатора, за счет использования лишь небольшого числа обучающих примеров, а также некоторых сторонних ресурсов, таких как параллельные корпуса, тезаурусы, базы знаний, анализаторы. Эти методы позволяют компенсировать малый размер вручную размеченных данных за счет статистической информации, содержащейся в большой неразмеченной выборке.

Настоящее исследование в первую очередь было ориентировано на создание методов и инструментов семантического анализа текстов на русском языке. Набор ресурсов, пригодных для обучения семантических анализаторов текстов на русском языке, сильно ограничен. Поэтому разработка методов определения ролевых структур высказываний на основе подходов с частичным привлечением учителя для русского языка является актуальным направлением исследований.

В настоящей работе предлагается использовать анализаторы на основе правил и словарей для автоматической семантической разметки крупных корпусов текстов, в которых ручная разметка семантических ролей отсутствует. Полученные таким образом корпуса можно использовать для обучения новых анализаторов для определения ролевых структур высказываний. Преимущество применения машинного обучения достигается за счет использования эталонной морфологической и синтаксической разметки при генерации ролей анализаторами на основе правил и словарей, тогда как обучение нового анализатора проводиться на автоматически сгенерированной морфологической и синтаксической разметке. В работе предлагаются также способы обучения на автоматически сгенерированной семантической разметке анализаторов для определения ролей при «неизвестных» предикатных словах, а также способ комбинации нескольких анализаторов на основе машинного обучения и словарного семантического анализатора для повышения качества определения ролевых структур высказываний. В ходе исследования была проведена ручная разметка аргументов, предикатных слов и ролей и создан тестовый корпус тек-

стов, на котором были проведены экспериментальные исследования разработанных анализаторов.

В первом разделе настоящей работы приведен обзор методов определения ролевых структур высказываний, основанных на машинном обучении с частичным привлечением учителя. Во втором разделе кратко описана модель текста, основанная на семантических ролях, а также словарный семантический анализатор, который используется для автоматической разметки обучающих корпусов и задает базовый уровень качества анализа при оценке новых методов. В третьем разделе описаны разработанные анализаторы на основе машинного обучения. В четвертом разделе представлены условия и результаты экспериментальных исследований разработанных анализаторов. В заключении приводятся результаты работы и предлагаются направления дальнейших исследований.

## 1. Обзор методов определения ролевых структур высказываний, основанных на машинном обучении с частичным привлечением учителя

Один из подходов к созданию семантических анализаторов, который можно отнести к обучению с частичным привлечением учителя, известен под названием «проекция аннотаций» (annotation projection). В нем семантическая структура размеченных предложений на одном языке сопоставляется с неразмеченными предложениями, переведенными на другой язык. Для использования подобного подхода необходимы параллельные корпуса для пары языков, для одного из которых существует семантический анализатор. Проекция аннотаций использовалась для создания из англоязычного корпуса FrameNet семантически размеченных корпусов на китайском [17], итальянском [18], немецком [19] и шведском [20] языках. Стоит также отметить работу [21], в которой параллельный корпус использовался для адаптации обученного анализатора к новым языкам путем сопоставления признаковых описаний аргументов из разных языков.

Другой подход заключается в обобщении фреймов (ролевых структур предикатных слов) на «неизвестные» предикатные слова (отсутствующие в словаре или в исходном обучающем корпусе). Для этого оценивают близость контекстов предикатных слов [22], проверяют наличие гипонимии между известными и неизвестными предикатными словами с помощью тезаурусов [23] (например, WordNet [24]), решают задачу классификации неизвестного предикатного слова по известным фреймам [25, 26].

Еще один подход заключается в создании новых размеченных примеров (предложений) на основе оценки их близости к размеченным примерам. В работе [27] на основе небольшого семантически размеченного корпуса автоматически строится большой размеченный корпус, который используется для обучения более точных семантических анализаторов. Новые аннотации строятся путем переноса ролевых структур предложений из размеченного корпуса на близкие к ним предложения из неразмеченного корпуса. Близость предложений вычисляется путем сопоставления их синтаксических деревьев и лексических признаков. В [28] предложен подход для адаптации семантической разметки к новой предметной области путем замены лексики в исходном корпусе на лексику из интересуемой предметной области. Для замены используется модель на основе рекуррентных нейронных сетей, WordNet и некоторые эвристики.

Отметим, что существует ряд подходов к решению задачи определения ролевых структур высказываний, основанных на машинном обучении без учителя (unsupervised machine learning), для которых не требуется предварительная семантическая разметка текста или какие-либо знания о семантических конструкциях естественного языка. Принцип работы таких методов заключается в кластеризации аргументов по набору признаков, извлеченных из их контекста. При этом алгоритм кластеризации или функционал качества задаются таким образом, чтобы порождаемые кластеры приближенно соответствовали семантическим ролям. Такие методы не способны самостоятельно именовать роли. Однако они могут быть использованы тогда, когда именование не важно, например, в поисковых машинах в Интернет или других системах, обрабатывающих большие объемы текстовых данных разных тематик и из разных предметных областей. Кроме того, многие из них могут быть модифицированы так, чтобы при формировании кластеров учитывался небольшой набор размеченных примеров, что позволяет строить более предсказуемые кластеры, близкие к необходимому результату.

В работе [29] описывается порождающая модель для определения ролевых структур высказываний без размеченного обучающего корпуса. В схеме порождения ролей отдельно учитывается порядок следования аргументов и вид аргумента. В работе [30] применяется метод агломеративной иерархической кластеризации для группировки аргументов по семантическим ролям (кластерам). В этом методе оценивается расстояние между кластерами по морфолексическим признакам входя-

щих в них аргументов с учетом того, что при одном предикатном слове не может быть двух аргументов с одинаковой ролью. Кластеры, расстояние между которыми меньше заданного порога, итеративно сливаются, при этом через каждые несколько итераций порог уменьшается для обеспечения сходимости. В работе [31] для решения задачи определения ролевых структур высказываний применен подход на основе минимизации ошибки восстановления входных данных. Предложенная модель состоит из двух компонент. Первая компонента (кодировщик) порождает роль при заданном аргументе и предикатном слове. Вторая компонента (декодировщик) порождает аргумент при заданной роли, предикатном слове и остальных аргументах. Кодировщик и декодировщик обучаются совместно так, чтобы минимизировать после этих преобразований ошибки восстановления входных данных. После обучения кодировщик используется для определения семантических ролей аргументов.

Аналитический обзор показывает, что исследования в области методов определения ролевых структур высказываний в текстах на русском языке, основанных на машинном обучении с частичным привлечением учителя, не проводились. Разработка подходов к определению ролевых структур высказываний, основанных на машинном обучении с частичным привлечением учителя особенно актуальна для создания семантических анализаторов русскоязычных текстов, поскольку известные ресурсы (FrameBank [4] и корпус, разрабатываемый в ИСА ФИЦ ИУ РАН [32]), в которых присутствует соответствующая разметка, пока еще недостаточно репрезентативны для эффективного применения методов машинного обучения с учителем. Несмотря на то, что обучение при помощи таких корпусов возможно [33], результат обученных анализаторов в реальных задачах не будет обладать достаточной полнотой. Например, в корпусе FrameBank только около 700 глаголов имеют размеченные примеры.

В современных работах мало внимания уделяется методам обучения на автоматически размеченных корпусах. Для русского языка было создано несколько семантических анализаторов на основе правил, решающих задачи близкие определению ролевых структур высказываний в текстах на русском языке: система для построения семантического графа А. Сокирко [34], словарный семантический анализатор, разработанный в ИСА ФИЦ ИУ РАН [32, 35, 36] и др. Они в основном опираются на вручную составленные семантические словари и большое количество эвристик, которые довольно трудно отлаживать и поддерживать. При этом они слабо устойчивы к ошибкам в морфологической и синтаксической разметке предложений текстов. Несмотря на это подобные анализаторы содержат в себе некоторое количество лингвистических знаний, которые могли бы быть полезны при создании систем семантического анализа.

## 2. Задача определения ролевых структур высказываний. Словарный семантический анализатор

Задача определения ролевых структур высказываний в текстах на естественном языке в разных работах ставится по-разному. В настоящей работе эта задача представлена следующим образом. В каждом предложении текста выделить предикатные слова, связанные с ними семантические аргументы и определить семантическую роль аргумента при предикатных словах. Предикатными словами могут быть глаголы, причастия, деепричастия, краткие прилагательные, отглагольные существительные, а также конструкции, состоящие из вспомогательного слова («можно», «нужно», «быть» и др.) и глагола или краткого прилагательного. Синонимичные предикатные слова сгруппированы в словарные статьи, схожим образом организованы фреймы во FrameNet. Семантическими аргументами в основном являются синтаксические вершины именных групп, управляемое слово в предложных группах (предлог может быть составным), числительные, а также некоторые наречия. Инвентарь ролей, используемый в настоящей работе, основан на реляционно-ситуационной модели текста [37], которая в свою очередь основана на теории коммуникативной грамматики Золотовой [38]. В него входит около 80 ролей, которые имеют общий смысл для всех предикатных слов (в отличие от ролей в корпусе Propbank, в котором один и тот же идентификатор роли может иметь отличный смысл при разных предикатных словах).

Для определения ролевых структур высказываний был разработан семантический анализатор, в основе которого лежит семантический словарь [32, 35, 36]. В словаре собрана информация о ролевых структурах предикатных слов и признаках, которыми должны обладать аргументы, чтобы им можно было назначить соответствующую семантическую роль. Семантический словарь состоит из словарных статей, в которых указаны: группа предикатных слов со схожим смыслом, набор ролей, которыми могут обладать аргументы при соответствующих предикатных словах, а также признаки, которыми должны обладать аргументы, чтобы получить ту или иную роль: падеж, предлог и категориально-семантический класс аргумента

(обобщенный категориальный смысл лексемы аргумента).

Рассмотрим алгоритм работы словарного семантического анализатора. Каждое предложение в анализаторе обрабатывается отдельно от остальных. На первом этапе выполняется поиск предикатных слов. С помощью набора эвристик выделяются глаголы, причастия, деепричастия, а также сложные конструкции, состоящие из нескольких слов, в которые, например, входят вспомогательные глаголы, краткие прилагательные, предикативы. Для каждого предикатного слова по лемме главного слова формируется список словарных статей, которым это предикатное слово может соответствовать.

На втором этапе для выделенных в предложении предикатных конструкций на основе анализа синтаксического дерева предложения осуществляется определение потенциальных семантических аргументов. На основе ряда эвристик аргументам назначается вес, а также определяются их характеристики.

На третьем этапе каждый выделенный потенциальный аргумент при заданном предикатном слове анализируется, и для каждой словарный статьи предикатного слова на основе сопоставления характеристик аргумента с признаками роли в словарной статье семантического словаря определяется набор ролей, которыми аргумент потенциально может обладать. Помимо этого, в анализаторе предусмотрен ряд эвристик, которые используются для назначения «необязательных» («периферийных» или «non-core») ролей. Для каждой роли также определяется ее вес.

На четвертом этапе проводится процедура разрешения неоднозначности распределения ролей по аргументам. В ней семантические роли распределяются между аргументами так, чтобы каждый семантический аргумент имел не более одной роли для заданного предикатного слова и словарной статьи, и каждая роль для заданной статьи использовалась бы не более одного раза. При этом покрытие аргументов ролями должно быть наилучшим в соответствии с весами аргументов и ролей. Это задача о назначениях, которая решается венгерским методом путем нахождения максимального паросочетания в двудольном графе. В результате решения этой задачи также вычисляется оценка словарной статьи предикатного слова. В итоге на основе этих оценок выбирается наилучшая словарная статья и соответствующий ей набор ролей. Алгоритм построения ролевой структуры предложения в словарном семантическом анализаторе представлен на рис. 2.

## 3. Методы семантического анализа текста, использующие машинное обучение на автоматически размеченном корпусе

### 3.1. Обучение на автоматически размеченном корпусе

В настоящем исследовании было предложено использовать анализаторы, основанные на правилах и словарях, для автоматической семантической разметки корпусов текстов. На полученных таким образом корпусах можно проводить машинное обучение новых анализаторов для определения ролевых структур высказываний. Однако проблема непосредственного применения такого подхода состоит в том, что качество новых анализаторов не может значимо превосходить качество исходного анализатора, с помощью которого был размечен корпус, поскольку его разметка используется для порождения обучающих примеров. Таким образом, простое использование анализаторов, обученных на автоматически сгенерированных данных, не дает преимуществ перед использованием анализаторов, основанных на правилах и словарях.

Однако обучение на автоматически размеченных корпусах может дать преимущество, если генерация семантических ролей (обучающих примеров) будет осуществляться на эталонной синтаксической и морфологической разметке с помощью словарного (основанного на правилах) анализатора, а обучение будет проводится на полученных ролевых структурах при наличии лишь автоматически сгенерированной морфосинтаксической разметки. Анализаторы, основанные на словарях и правилах, на эталонной морфосинтаксической разметке обычно работают значительно лучше, чем на разметке, сгенерированной автоматически. Тогда как анализаторы на основе машинного обучения часто могут автоматически адаптироваться к ошибкам нижележащих уровней разметки за счет использования большого признакового пространства и выдавать более устойчивые к ошибкам результаты. Таким образом, если имеется значительная разница в качестве работы анализатора, генерирующего обучающие примеры, на эталонной морфосинтаксической разметке и на автоматически сгенерированной разметке, то применение машинного обучения может помочь сократить эту разницу.

На момент проведения исследования наиболее крупным русскоязычным корпусом, содержащим эталонную морфологическую и синтаксическую разметку, является корпус СинТагРус [39]. Версия корпуса, которая использовалась в исследовании, содержит 53 439 предложений и 774 373 токенов без учета пунктуации. Корпус не содержит

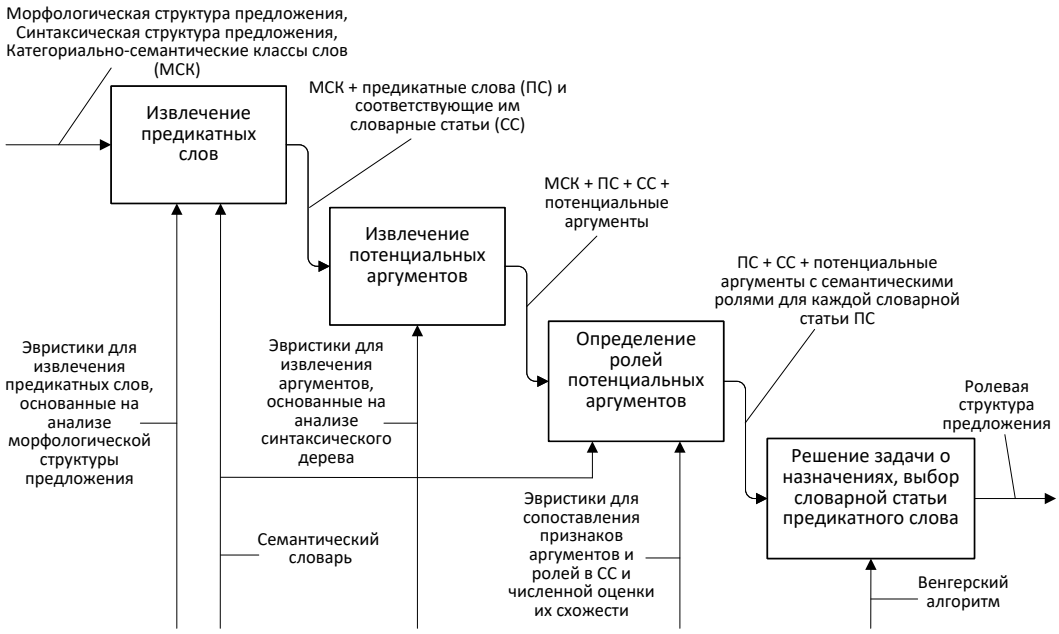


**Рис. 2.** Диаграмма построения ролевой структуры предложения в словарном семантическом анализаторе

разметку ролевых структур высказываний, поэтому не может быть использован напрямую для обучения необходимого семантического анализатора. Было предложено провести автоматическую семантическую разметку корпуса СинТагРус с помощью словарного семантического анализатора, разработанного в ИСА ФИЦ ИУ РАН. Предварительные эксперименты показали, что словарный анализатор на эталонной морфологической и синтаксической разметке корпуса СинТагРус генерирует ролевые структуры с существенно меньшим числом ошибок, чем на разметке, сгенерированной автоматически. Полученная таким образом разметка семантических ролей, аргументов и предикатных слов была использована для создания семантического анализатора на основе машинного обучения. При этом для генерации признаков использовалась не эталонная синтаксическая и морфологическая разметка, а разметка, полученная автоматически с помощью АОТ [40] и MaltParser [41] соответственно. На рис. 3 представлена схема процесса обучения анализатора на автоматически сгенерированном семантическом корпусе.

В ходе исследования был предложен ряд моделей машинного обучения, основанных на разных подходах к решению задач выявления аргументов предикатных слов и непосредственно назначения метки-роли аргументу. Предварительные эксперименты показали, что использование модели на основе машинного обучения непосредственно для определения семантической роли аргумента независимо от остальных аргументов дает преимущество по сравнению с определением роли по семантическому словарю.

Алгоритм использования такой модели следующий. Вместо применения эвристик для сопоставления признаков потенциальных аргументов и признаков ролей в семантическом словаре, каждый потенциальный аргумент подается на вход классификатору, который определяет степень уверенности в принадлежности аргумента каждому из классов, обозначающих семантическую роль (около 80 классов) – вес роли. В зависимости от используемого метода классификации вес роли представляют собой либо аппроксимированную вероятность роли по обученной модели, либо значение решающей функции классификатора. Семантические роли с низким весом отсекаются по порогу. В результате формируется набор потенциальных семантических аргументов, которым сопоставлено несколько ролей с весами (в том числе аргументу может быть не сопоставлено ни одной роли). Полученный набор аргументов и ролей обрабатывается с помощью венгерского алгоритма, который распределяет роли по аргументам так, чтобы каждый семантический аргумент имел не более одной роли для заданного предикатного слова и словарной статьи, и каждая роль для заданной статьи использовалась бы не более одного раза. Это выполняется для каждой словарной статьи предикатного слова, после чего выбирается словарная статья и набор ролей, соответствующие словарной статье с

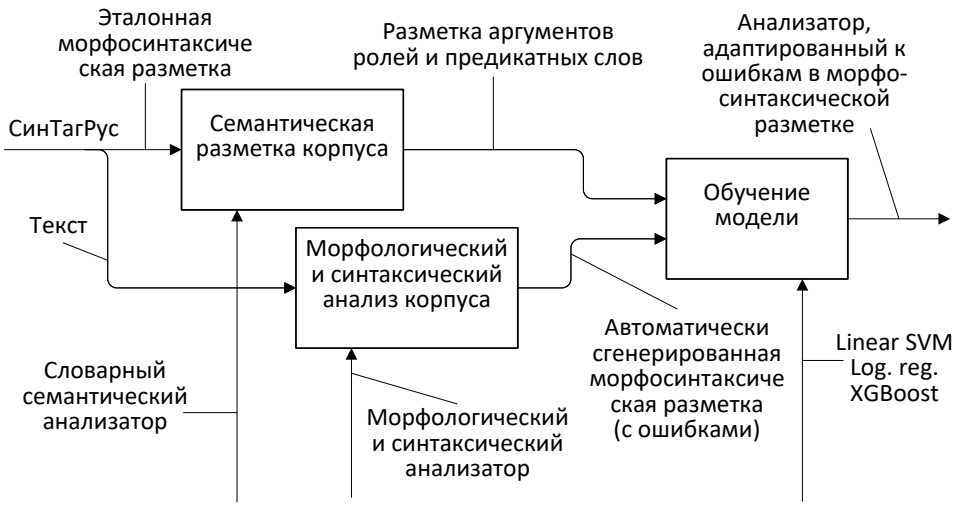


**Рис. 3.** Диаграмма обучения анализатора для определения ролевых структур высказываний на автоматически сгенерированном корпусе

максимальной оценкой, полученной в результате решения задачи о назначениях. На рис. 4 представлен процесс семантического анализа с применением машинного обучения.

В качестве признаков для классификации использовались различные синтаксические и лексико-морфологические характеристики аргументов, предикатных слов и структура предложения, а также их комбинации. Рассмотрим основные признаки модели для определения роли потенциального аргумента:

- падеж главного слова аргумента;
- предлог аргумента;
- категориально-семантический класс аргумента;
- лемма аргумента;
- идентификатор словарной статьи предикатного слова;
- лемма аргумента + идентификатора словарной статьи предикатного слова;
- падеж + предлог аргумента;
- идентификатор словарной статьи предикатного слова + падеж + предлог аргумента;
- падеж + предлог + категориально-семантический класс аргумента;
- части речи предикатного слова + падеж аргумента;
- часть речи предикатного слова + падеж + предлог аргумента;
- время предикатного слова;
- возвратность предикатного слова;
- форма предикатного слова (глагол, причастие, деепричастие);
- позиция аргумента относительно предикатного слова;
- является ли аргумент вопросительным словом («что», «когда», «где» и др.);
- глубина аргумента в синтаксическом дереве по отношению к предикатному слову.

Признаковое пространство классификатора имеет большой размер и является разреженным. В представленной модели присутствует более 120 000 признаков. Из-за этой особенности для обеспечения приемлемой скорости обработки в качестве метода классификации в основном исследовались линейные модели. По итогам предварительных экспериментов в качестве метода машинного обучения был выбран метод опорных векторов с линейным ядром.

### 3.2. Определения ролей аргументов при «неизвестных» предикатных словах

В работе предложено использовать машинное обучение также и для определения ролей аргументов при «неизвестных» предикатных словах (отсутствующих в словаре). Словарный семантический анализатор по сути не обрабатывает аргументы для предикатных слов, для которых отсутствует словарная статья, поскольку отсутствуют необходимые признаки для сопоставления, по которым можно было бы выбрать семантические роли аргументов.

Машинное обучение частично помогает решить эту проблему. Хотя роли аргументов сильно зависят от предикатного слова, обозначающего ситуацию, аргументы, имеющие схожие контекстные признаки, независящие непосредственно от предикатных слов, также имеют и схожие роли. Поэтому можно ожидать, что в задаче классификации аргумента по ролям в случае «неизвестного» предикатного слова, отсутствие информации о преди-

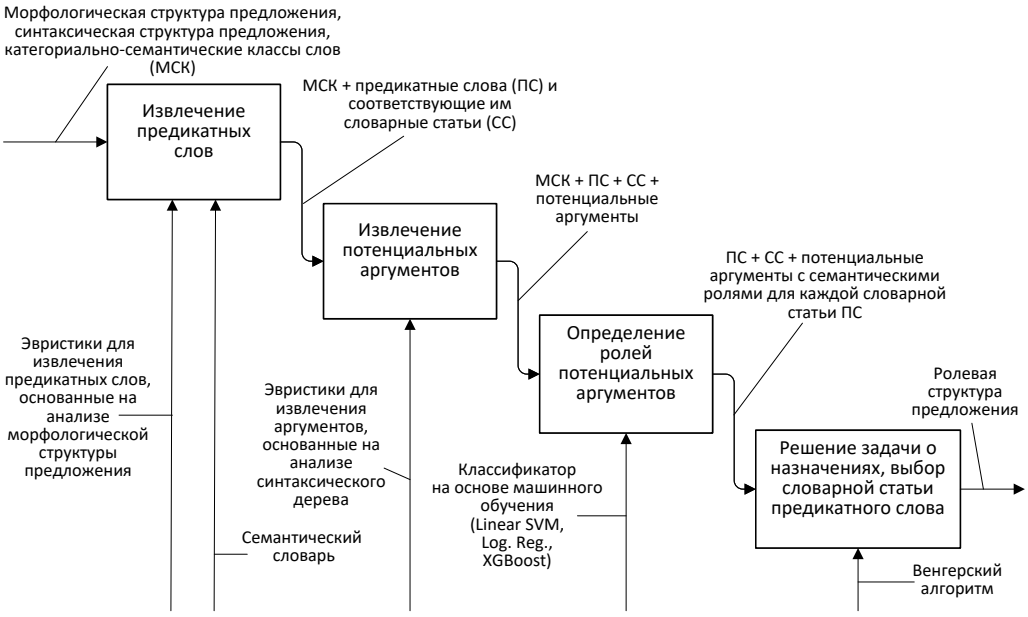


**Рис. 4.** Диаграмма построения ролевой структуры предложения в семантическом анализаторе, использующем машинное обучение

катном слове можно частично компенсировать за счет широкого набора признаков, извлеченных из контекста аргумента и предикатного слова.

Для определения ролей аргументов при «неизвестных» предикатных словах на корпусе СинТагРус, автоматически размеченном словарным семантическим анализатором, была обучена модель без признаков, учитывающих идентификаторы словарных статей. Это позволяет проводить обобщение между различными словарными статьями и назначать роли аргументам при «неизвестных» предикатных словах. Таким образом, признаки модели для определения ролей «неизвестных» предикатных слов включают все признаки модели для назначения метки-роли, кроме тех, в которых фигурирует идентификатор словарной статьи, плюс также следующие признаки:

- категориально-семантический класс + падеж + предлог + позиция аргумента относительно предикатного слова;
- падеж + предлог + форма предикатного слова (форма глагола);
- падеж + предлог + форма предикатного слова (форма глагола) + флаг «возвратность» глагола;
- падеж + предлог + форма предикатного слова (форма глагола) + флаг «возвратность» + категориально-семантический класс аргумента;
- падеж + предлог + форма предикатного слова (форма глагола) + флаг «возвратность» + категориально-семантический класс + позиция аргумента относительно предикатного слова.

Исследовалась также возможность замены идентификатора словарной статьи на векторное представление леммы предикатного слова (word embedding) [42]. Во многих работах было показано, что подобные представления хорошо описывают смысловую близость слов [43]. Поэтому в случае, если «неизвестное» предикатное слово близко по смыслу с «известным» предикатным словом, то ролевую структуру «неизвестного» предикатного слова потенциально можно предсказать на основе близости их векторных представлений лемм. В альтернативный набор признаков для задачи определения ролей при «неизвестных» предикатных словах дополнительно были включены векторные представления лемм аргументов, векторные представления лемм предикатных слов, а также декартово произведение векторного представления леммы предикатного слова и вектора признаков для конкатенации «падеж + предлог + категориально-семантический класс». Векторные представления были получены из ресурса RusVectores [44], использовалась модель, построенная на основе НКРЯ.

В качестве метода классификации в анализаторе для определения ролей аргументов при «неизвестных» предикатных словах также использовался метод опорных векторов с линейным ядром. В итоге был создан «составной» анализатор, в котором используются две модели машинного обучения: первая определяет роли аргументов в случае, когда предикатное слово присутствует в семантическом словаре (список выделенных для него словарных

статей не пуст), а вторая определяет роли, когда предикатное слово в словаре найдено не было.

### 3.3. Гибридный семантический анализатор

Как известно, применение моделей, обученных на данных из одной предметной области, к сильно отличающимся данным из другой предметной области может привести к значительному падению качества классификации. Естественным является тот факт, что модели не всегда могут уверенно работать на объектах, которые сильно отличаются от обучающих примеров.

В случае с обучением на корпусе СинТагРус с автоматически сгенерированной семантической разметкой стоит отметить, что корпус содержит не все предикатные слова из семантического словаря и, следовательно, в обучающей выборке отсутствуют необходимые примеры для правильной классификации аргументов по ролям в таких случаях с помощью модели для «известных» предикатных слов. При работе модели это проявляется в том, что все классы (роли) имеют одинаково низкие оценки, ниже чем минимальный порог отсечения. Это происходит из-за того, что большое количество важных признаков являются составными и опираются на лемму или идентификатор словарной статьи предикатного слова. Если идентификатор и лемма в обучающем корпусе не встречаются, эти признаки имеют нулевое значение в ходе классификации аргумента, и значение решающей функции будет небольшим. Однако при этом предикатное слово может содержаться в семантическом словаре, и, таким образом, словарный семантический анализатор будет обрабатывать его корректно за счет заложенных в нем эвристик. Таким образом, эта проблема может быть решена тремя способами:

- Расширение обучающего корпуса за счет автоматического семантического анализа текстов без эталонной морфосинтаксической разметки. Это позволит увеличить покрытие предикатных слов в семантическом словаре, однако ошибки в морфосинтаксической разметке могут сильно ухудшить качество анализа и нивелировать преимущества модели, обученной на ролях, полученных на эталонной разметке.
- Применение модели для определения ролей «неизвестных» предикатных слов.
- Использование словарного семантического анализатора в тех случаях, когда модель машинного обучения не может уверенно предсказать ни одной роли.

В настоящем исследовании был реализован третий способ. В итоге был создан гибридный семантический анализатор, который осуществляет определение ролевых структур высказываний с помощью классификаторов на основе машинного обучения и с помощью эвристик по семантическому словарю.

Алгоритм его работы следующий. После извлечения предикатных слов и потенциальных семантических аргументов, для каждого предикатного слова проводится поиск подходящих для него словарных статей. Если хотя бы одна словарная статья найдена, то для определения ролей аргументов используется модель «известных» предикатных слов. Если после работы этой модели ни одна роль не получает оценку выше заданного порога, то они определяются с помощью словарного семантического анализатора. Если ни одной словарной статьи найдено не было, то для определения ролей аргументов используется модель для «неизвестных» предикатных слов. На рис. 5 представлен алгоритм работы гибридного семантического анализатора.

## 4. Экспериментальные исследования разработанных методов

### 4.1. Тестовый и обучающий корпус, метрики оценки качества

Для экспериментальных исследований разработанных анализаторов был размечен тестовый подкорпус СинТагРус, состоящий из 250 предложений. Он является частью семантически размеченного корпуса размером 1500 предложений, разработанного в ИСА ФИЦ ИУ РАН ранее [32]. Недостатком последнего являлось отсутствие разметки необязательных (периферийных) ролей, низкое покрытие деепричастий и причастий, а также отсутствие разметки для предикатных слов, не представленных в семантическом словаре. В новом тестовом корпусе эти недостатки были устранены: в нем были размечены все предикатные слова (глаголы, причастия, деепричастия, краткие прилагательные), аргументы и их семантические роли. Исключением являются предикатные слова-существительные, их покрытие обуславливается семантическим словарем. В тестовом корпусе содержится около 1200 семантических аргументов, около 600 предикатных слов и 54 различные роли.

Для обучения использовалась часть СинТаг Рус, в которой содержится около 48 000 предложений и 700 000 токенов без учета пунктуации. Обучающий корпус не пересекается с тестовым. В результате разметки словарным семантическим анализатором в обучающем корпусе было выделено около 81 000 предикатных слов, 123 000 аргументов и 84 различных ролей.

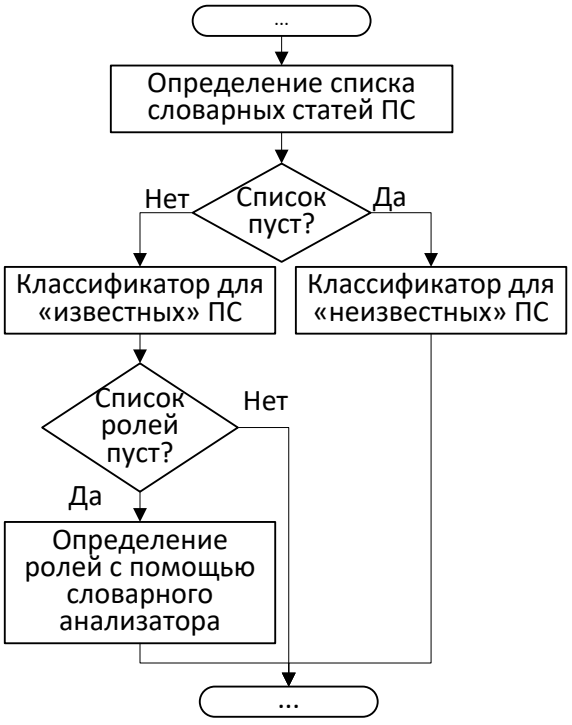


**Рис. 5.** Алгоритм работы гибридного семантического анализатора

В качестве оценки качества работы семантических анализаторов использовались стандартные метрики: точность *p* (precision), полнота *r* (recall) и $F_1$-мера. При сопоставлении результатов анализаторов с тестовым корпусом учитывались только те семантические роли и аргументы, которые присутствуют в тестовом корпусе. Для оценки качества классификации также использовалась доля правильных ответов *Acc* (accuracy).

### 4.2. Эксперименты и результаты

В первую очередь была проведена оценка различий в качестве работы словарного семантического анализатора на эталонной морфосинтаксической разметке и на автоматически сгенерированной разметке. В табл. 1 представлены результаты оценки на тестовом корпусе:

- Dict_clean_annots – качество работы на эталонной морфосинтаксической разметке;
- Dict_dirty_annots – качество работы на автоматически сгенерированной морфосинтаксической разметке.

**Таблица 1**

Сравнение качества работы словарного семантического анализатора на эталонной морфосинтаксической разметке

| Анализатор | *p*,% | *r*,% | $F_1$,% |
|---|---|---|---|
| Dict_clean_annots | 85,6 | 55,1 | 67,0 |
| Dict_dirty_annots | 84,6 | 48,1 | 61,3 |

Из проведенного эксперимента видно, что влияние на качество семантического анализа ошибок в морфосинтаксической разметке составляет около $\Delta F$=5,3%; $\Delta r$=7,0%, что является весьма значительным. Таким образом, в данной ситуации имеет смысл применять методы машинного обучения, которые могут сократить эту разницу. Стоит отметить, что полнота анализатора несколько ниже, чем результаты, представленные в [35]. Это частично вызвано отличием способа разметки однородных семантических аргументов (анализатор указывает роль только для первого из однородных членов, тогда как в тестовом корпусе размечены все однородные члены), а также тем, что в тестовом корпусе, который используется в настоящей работе, размечены все аргументы, в том числе и при «неизвестных» предикатных словах.

Были проведены экспериментальные исследования разработанных анализаторов для определения ролевых структур высказываний, использующих машинное обучения на автоматически размеченном корпусе. В табл. 2 представлены результаты и в введены следующие обозначения:

- Dict_dirty_annots – результат работы словарного анализатора на автоматически сгенерированной синтаксической и морфологической разметке;
- ML – результат работы анализатора на основе машинного обучения на автоматически сгенерированной синтаксической и морфологической разметке;
- Comb_ML – результат работы составного анализатора, в котором используются две модели машинного обучения для «известных» и «неизвестных» предикатных слов;
- Comb_ML_embeddings – результат работы составного анализатора, в котором помимо стандартных признаков используются признаки на основе векторных представлений лемм;
- Hybrid_Comb_ML_embeddings – гибридный составной обучаемый анализатор с признаками на основе векторных представлений лемм, в котором помимо классификатора используется также словарный анализатор в случаях, когда основной анализатор выдает низкие веса ролей.

**Таблица 2**

Результаты экспериментальных исследований семантических анализаторов

| **Анализатор** | ***p*,%** | ***r*,%** | $F_1$**,%** |
|---|---|---|---|
| Dict_dirty_ annots | 84,6 | 48,1 | 61,3 |
| ML | 84,9 | 49,7 | 62,7 |
| Comb_ML | 84,5 | 51,8 | 64,2 |
| Comb_ML_ embeddings | 84,5 | 51,9 | 64,3 |
| Hybrid_Comb_ ML_ embeddings | 84,1 | 52,8 | 64,8 |

Была проведена также оценка качества работы непосредственно самих классификаторов. Для этого вычислялась макроусредненная доля правильных ответов при помощи пятипроходной перекрестной проверки на обучающем корпусе. Результаты представлены в табл. 3. Оценивались:

- ML – стандартная модель для определения роли аргумента при «известном» предикатном слове;
- Unknown_pred – модель для определения роли аргумента при «неизвестном» предикатном слове с помощью набора признаков без вложений;
- Unkown_pred_embeddings – модель для определения роли аргумента при «неизвестном» предикатном слове с помощью набора признаков с векторными представлениями лемм.

**Таблица 3**

Результаты оценки качества классификаторов с помощью перекрестной проверки на обучающем корпусе

| **Анализатор** | ***Acc.*, %** |
|---|---|
| ML | 94,4 |
| Unknown_pred | 69,2 |
| Unkown_pred_ embeddings | 83,9 |

Результаты показывают, что использование обучения на автоматически сгенерированном корпусе позволяет сократить разрыв между качеством работы словарного анализатора на автоматически сгенерированной морфологической и синтаксической разметке и эталонной разметке за счет использования машинного обучения. Обучаемый анализатор имеет большую полноту ($\Delta p$=1,6 %) при схожей точности, что в итоге дает прирост $F_1$-меры ($\Delta F_1$=1,4 %). Это достигается за счет того, что в отличие от словарного анализатора классификатор на основе машинного обучения позволяет адекватно определять роли даже в тех случаях, когда в значениях некоторых из признаков содержится шум. Стоит также отметить, что разница в качестве работы словарного анализатора частично обусловлена ошибками при определении семантических аргументов, которые не могут быть исправлены моделью на основе машинного обучения, поскольку алгоритм работы анализатора предполагает, что модель классифицирует только те аргументы, которые уже были выделены с помощью эвристик. Из табл. 3 видно, что классификатор совершает мало ошибок при перекрестной проверке на обучающих данных, и, таким образом, представленный набор признаков позволяет достаточно хорошо решать поставленную задачу определения семантической роли для заданного аргумента.

Использование составного анализатора, в котором применяются две модели машинного обучения для «известных» и «неизвестных» предикатных слов, весьма существенно повышает полноту определения ролевых структур высказываний, учитывая небольшую долю «неизвестных предикатных слов» среди всех предикатных слов в тестовом корпусе (7-8%). При этом ожидаемо снижается точность из-за новых ошибок на аргументах «неизвестных» предикатных слов. При использовании стандартных признаков прирост полноты составляет $\Delta p$=2,1 %, что при уменьшении точности дает также и прирост $\Delta F_1$=1,5 %. Использование вложений в задаче определения ролей для аргументов «неизвестных» предикатных слов предположительно дает существенное преимущество. Хотя относительный прирост качества решения задачи определения ролевых структур высказываний за счет использования вложений – небольшой, среди аргументов «неизвестных» предикатных слов он может оказаться весьма значимым. Из табл. 3 видно, что добавление векторных представлений вместо идентификаторов предикатных слов в значительной степени нивелирует разницу между результатами классификаторов для «известных» и «неизвестных» предикатных слов со стандартными признаками. Однако это не может считаться показательным, поскольку при перекрестной проверке в тестовую выборку подмешиваются примеры, предикатные слова которых присутствуют и в обучающей выборке. Таким образом, для определения эффекта от использования в признаковом

наборе векторных представлений «неизвестных» предикатных слов необходимо больше соответствующих тестовых примеров и более репрезентативный тестовый корпус. Использование векторных представлений дает прирост качества анализа относительно использования стандартного набора признаков: $\Delta p$=2,2 %, а $\Delta F_1$=1,6 %. Можно сделать вывод о том, что применение дополнительной модели машинного обучения для «неизвестных» предикатных слов положительно сказывается на качестве работы анализатора в целом.

Использование словарного семантического анализатора совместно с двумя моделями машинного обучения позволяет добиться наилучших результатов. Гибридный семантический анализатор по сравнению с анализатором только лишь на основе машинного обучения также дает прирост полноты $\Delta p$=0,9% при $\Delta F_1$=0,5%. По отношению к исходному словарному анализатору гибридный семантический анализатор дает значительный прирост полноты и $F_1$-меры ($\Delta p$=4,7 %, $\Delta F_1$=3,5 %).

## Заключение

В работе предложены и исследованы подходы к определению ролевых структур высказываний, использующие принципы машинного обучения с частичным привлечением учителя. Рассмотрено обучение на автоматически размеченном корпусе с помощью словарного семантического анализатора. Показано, что за счет машинного обучения можно уменьшить число ошибок в семантическом анализе, если для генерации обучающей выборки для задачи определения ролевых структур высказываний использовать корпус с эталонной морфосинтаксической разметкой. Из экспериментальных исследований видно, что представленные в работе признаки позволяют проводить классификацию с высокой точностью.

Предложен подход к определению ролевых структур высказываний для «неизвестных» предикатных слов, которые отсутствуют в семантическом словаре словарного анализатора. За счет использования большого признакового пространства, включая векторные представления лемм, во многих случаях удается корректно определять роли семантических аргументов при «неизвестных» предикатных словах, за счет чего повышается полнота и качество семантического анализа в целом.

В работе также представлен гибридный подход к семантическому анализу, в котором используются две модели машинного обучения для «известных» и «неизвестных» предикатных слов, а также словарный семантический анализатор. Экспериментальные исследования показывают, что за счет использования словарного анализатора для обработки случаев, в которых классификаторы не могут с высокой уверенностью определить роль, также можно повысить полноту и общее качество семантического анализа.

Среди перспективных подходов к повышению качества семантического анализа можно назвать подходы на основе самообучения. Они также позволяют повысить полноту анализа за счет итеративной доразметки обучающей выборки с помощью анализатора, обученного на корпусе из предыдущей итерации. Это в свою очередь уменьшает количество неразмеченных примеров словарным анализатором, что в итоге упрощает процесс обучения и позволяет получать более качественные модели для определения ролевых структур высказываний.

Применение активного обучения также может существенно повысить качество семантического анализа. Исследования в этом направлении могут помочь уменьшить долю ручного труда при создании больших семантически размеченных корпусов.

Методы определения ролевых структур высказываний на основе машинного обучения без учителя являются альтернативой стандартным подходам. В ряде работ показано, что в результате кластеризации можно выделить группы аргументов, которые приближенно соответствуют семантическим ролям. Использование подобных методов совместно с небольшой размеченной выборкой может, с одной стороны, помочь повысить их качество, а, с другой стороны, частично решить проблему малого количества обучающих данных.

Еще одним перспективным направлением исследований является смешивание результатов нескольких словарных (на основе) правил анализаторов в едином корпусе. Это позволит преобразовывать знания, заложенные в разных анализаторах, в форму (размеченный корпус), удобную для использования методами на основе машинного обучения.

## Литература

1 *Fillmore C. J.* The case for case // Universals in Linguistic Theory / Ed. by Emmon Bach, Robert T. Harms. — New York, 1968. — P. 1–88.

2 *Gildea D., Jurafsky D.* Automatic labeling of semantic roles // Computational Linguistics. — 2002. — Vol. 28, no. 3. — P. 245–288.

3 *Плунгян В. А.* Введение в грамматическую семантику: грамматические значения и грамматические системы языков мира: учебное пособие. М.: Издательство РГГУ, 2011.

4 *Кашкин Е. В., Ляшевская О. Н.* Семантические роли и сеть конструкций в системе FrameBank // Труды международной конференции «Диалог 2013». — 2013. — С. 325–343.
5 *Shen D., Lapata M.* Using semantic roles to improve question answering // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). — Association for Computational Linguistics, 2007. — P. 12–21.
6 *Kaisser M., Webber B.* Question answering based on semantic roles // Proceedings of the Workshop on Deep Linguistic Processing. — Association for Computational Linguistics, 2007. — P. 41–48.
7 *Семантико-синтаксический* анализ текстов в задачах вопросно-ответного поиска и извлечения определений / А.О. Шелманов, М.И. Каменская, И.В. Ананьева, И.В. Смирнов // Искусственный интеллект и принятие решений. — 2016. — № 4.
8 *Liu D., Gildea D.* Semantic role features for machine translation // Proceedings of the 23rd International Conference on Computational Linguistics. — Association for Computational Linguistics, 2010. — P. 716–724.
9 Relation alignment for textual entailment recognition / Mark Sammons, VG Vinod Vydiswaran, Tim Vieira et al. // Text Analysis Conference (TAC). — 2009.
10 *Xue N., Palmer M.* Calibrating features for semantic role labeling // Proceedings of EMNLP 2004. — Association for Computational Linguistics, 2004. — P. 88–94.
11 Shallow semantic parsing using support vector machines / Sameer S Pradhan, Wayne H Ward, Kadri Hacioglu et al. // HLT-NAACL 2004: Main Proceedings. — Association for Computational Linguistics, 2004. — P. 233–240.
12 *Toutanova K., Haghighi A., Manning C. D.* Joint learning improves semantic role labeling // Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. — Association for Computational Linguistics, 2005. — P. 589–596.
13 *Punyakanok V., Roth D., Yih W.-t.* The importance of syntactic parsing and inference in semantic role labeling // Computational Linguistics. — 2008. — Vol. 34, no. 2. — P. 257–287.
14 *Palmer M., Gildea D., Kingsbury P.* The proposition bank: An annotated corpus of semantic roles // Computational linguistics. — 2005. — Vol. 31, no. 1. — P. 71–106.
15 *Fillmore C. J., Johnson C. R., Petruck M. R.* Background to FrameNet // International journal of lexicography. — 2003. — Vol. 16, no. 3. — P. 235–250.
16 *The CoNLL-2009 shared task*: Syntactic and semantic dependencies in multiple languages / Jan Hajic, Massimiliano Ciaramita, Richard Johansson et al. // Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task. — Association for Computational Linguistics, 2009. — P. 1–18.
17 *Fung P., Chen B.* BiFrameNet: bilingual frame semantics resource construction by cross-lingual induction // Proceedings of the 20th international conference on Computational Linguistics. — Association for Computational Linguistics, 2004.
18 Cross-language frame semantics transfer in bilingual corpora / Roberto Basili, Diego De Cao, Danilo Croce et al. // International Conference on Intelligent Text Processing and Computational Linguistics / Springer. — 2009. — P. 332–345.
19 *Padó S., Lapata M.* Cross-lingual annotation projection for semantic roles // Journal of Artificial Intelligence Research. — 2009. — Vol. 36. — P. 307–340.
20 *Johansson R., Nugues P.* A FrameNet-based semantic role labeler for Swedish // Proceedings of the COLING/ACL. — Association for Computational Linguistics, 2006. — P. 436–443.
21 *Kozhevnikov M., Titov I.* Cross-lingual transfer of semantic role labeling models // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Association for Computational Linguistics, 2013. — P. 1190–1200.
22 *Das D., Smith N. A.* Semi-supervised frame-semantic parsing for unknown predicates // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. — Association for Computational Linguistics, 2011. — P. 1435–1444.
23 *Burchardt A., Erk K., Frank A.* A WordNet detour to FrameNet // Sprachtechnologie, mobile Kommunikation und linguistische Resourcen. — 2005. — Vol. 8. — P. 408–421.
24 *Miller G. A.* WordNet: A lexical database for English // Communications of the ACM. — 1995. — Vol. 38, no. 1. — P. 39–41.
25 *Johansson R., Nugues P.* Using WordNet to extend FrameNet coverage // In Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages at the 16th Nordic Conference of Computational Linguistics (NODALIDA). — 2007. — P. 27–30.
26 *Automatic induction* of FrameNet lexical units / Marco Pennacchiotti, Diego De Cao,

Roberto Basili et al. // Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. — Association for Computational Linguistics, 2008.

27 *Fürstenau H., Lapata M.* Semi-supervised semantic role labeling via structural alignment // Computational Linguistics. — 2012. — Vol. 38, no. 1. — P. 135–171.

28 *Do Q. T. N., Bethard S., Moens M.-F.* Domain adaptation in semantic role labeling using a neural language model and linguistic resources // IEEE/ACM Transactions on Audio, Speech, and Language Processing. — 2015. — Vol. 23, no. 11. — P. 1812–1823.

29 *Garg N., Henderson J.* Unsupervised semantic role induction with global role ordering // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. — Association for Computational Linguistics, 2012. — P. 145–149.

30 *Lang J., Lapata M.* Similarity-driven semantic role induction via graph partitioning // Computational linguistics. — 2014. — Vol. 40, no. 3. — P. 633–669.

31 *Titov I., Khoddam E.* Unsupervised induction of semantic roles within a reconstruction error minimization framework // In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2015.

32 *Shelmanov A. O., Smirnov I. V.* Methods for semantic role labeling of Russian texts // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (2014). — No. 13. — 2014. — P. 607–620.

33 *Kuznetsov I.* Semantic role labeling for Russian language based on Russian FrameBank // International Conference on Analysis of Images, Social Networks and Texts / Springer. — 2015. — P. 333–338.

34 *Сокирко А. В.* Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ) : Дисс кандидата наук / А. В. Сокирко. — 2001.

35 *Осипов Г. С., Шелманов А. О.* Метод повышения качества синтаксического анализа на основе взаимодействия синтаксических и семантических правил // Труды шестой международной конференции «Системный анализ и информационные технологии» (САИТ). Т. 1. — 2015. — С. 229–240.

36 *Семантико-синтаксический анализ* естественных языков Часть II. Метод семантико-синтаксического анализа текстов / И. В. Смирнов, А. О. Шелманов, Е. С. Кузнецова, И. В. Храмоин // Искусственный интеллект и принятие решений. — № 1. — С. 11–24.

37 *Осипов Г. С., Смирнов И. В., Тихомиров И. А.* Реляционно-ситуационный метод поиска и анализа текстов и его приложения // Искусственный интеллект и принятие решений. — 2008. — № 2. — С. 3–10.

38 *Золотова Г. А., Онипенко Н. К., Сидорова М. Ю.* Коммуникативная грамматика русского языка // Институт русского языка РАН им. В. В. Виноградова. — 2004.

39 *Синтаксически и семантически* аннотированный корпус русского языка: современное состояние и перспективы / Ю. Д. Апресян, И. М. Богуславский, Б. Л. Иомдин и др. // Национальный корпус русского языка: 2003–2005. — 2005. — С. 193–214.

40 *Автоматическая обработка текста*. — 2016. — окт. — URL: http://www.aot.ru/.

41 *MaltParser*: A language-independent system for data-driven dependency parsing / Joakim Nivre, Johan Hall, Jens Nilsson et al. // Natural Language Engineering. — 2007. — Vol. 13, no. 2. — P. 95–135.

42 *Distributed representations* of words and phrases and their compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // Advances in neural information processing systems. — 2013. — P. 3111–3119.

43 *Mnih A., Kavukcuoglu K.* Learning word embeddings efficiently with noise-contrastive estimation // Advances in Neural Information Processing Systems. — 2013. — P. 2265–2273.

44 *Kutuzov A., Andreev I.* Texts in, meaning out: neural language models in semantic similarity task for Russian // Proceedings of the Dialog Conference. — 2015.

**Шелманов Артем Олегович.** Младший науучный сотрудник ИСА ФИЦ ИУ РАН. Кандидат технических наук. Окончил в 2011 г. МИФИ. Количество печатных работ: 20. Область научных интересов: искусственный интеллект, компьютерная лингвистика, машинное обучение, информационно-аналитические системы. E-mail: shelmanov@isa.ru

**Каменская Маргарита Александровна.** Инженер-исследователь ИСА ФИЦ ИУ РАН. Окончила в 2014 г. РУДН. Количество печатных работ: 5. Область научных интересов: компьютерная лингвистика, обработка естественного языка, разрешение референции. E-mail: mak@isa.ru

## Training semantic role labeler for Russian using automatically annotated corpus

*A.O. Shelmanov, M.A. Kamenskaya*

**Abstract.** The paper describes the research of methods for semantic role labeling based on semi-supervised machine learning. We present a method for training semantic role labeler using corpus automatically annotated by baseline dictionary-based (rule-based) semantic parser that improves the performance of the baseline. We also propose a method for labeling arguments of "unknown" predicates that are not present in the semantic dictionary of the baseline parser. The hybrid semantic parser is presented. It uses two models for "known" and "unknown" predicates as well as the dictionary-based parser. The experiments with the manually labeled test corpus in Russian show that modifications proposed in the paper improve recall and overall performance of semantic role labeling.
**Keywords:** *semantic role labeling, semi-supervised machine learning, semantic parsing, word embedding.*

### References

1. *Fillmore C. J.* The case for case // Universals in Linguistic Theory / Ed. by Emmon Bach, Robert T. Harms. — New York, 1968. — P. 1–88.
2. *Gildea D., Jurafsky D.* Automatic labeling of semantic roles // Computational Linguistics. — 2002. — Vol. 28, no. 3. — P. 245–288.
3. Plungyan, V. A. 2011. Vvedenie v grammaticheskuyu semantiku: grammaticheskie znacheniya i gramma-ticheskie sistemy yazykov mira: uchebnoe posobie [Introduction to grammatical semantics: grammatical meanings and grammatical system of the world's languages]. Moscow: RSUH Publs. 672 p.
4. *Kashkin, E. V., Lyashevskaya, O. N.* 2013. Semanticheskie roli i set' konstruktsiy v sisteme FrameBank [Semantic roles and constructs network in FrameBank system]. Trudy mezhdunarodnoy konferentsii "Dialog 2013" [International Conference "Dialogue-2013"]. Moscow. 325–343.
5. *Shen D., Lapata M.* Using semantic roles to improve question answering // Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). — Association for Computational Linguistics, 2007. — P. 12–21.
6. *Kaisser M., Webber B.* Question answering based on semantic roles // Proceedings of the Workshop on Deep Linguistic Processing. — Association for Computational Linguistics, 2007. — P. 41–48.
7. *Shelmanov, A.O., Kamenskaya, M.A., Anan'eva, M.I., Smirnov, I.V.* 2016. Semantiko-sintaksicheskiy analiz tekstov v zadachakh voprosno-otvetnogo poiska i izvlecheniya opredeleniy [Semantic-syntactic analysis for question-answering and definition extraction]. Iskusstvennyy intellekt i prinyatie resheniy [Artificial intelligence and decision-making]. (In the press.)
8. *Liu D., Gildea D.* Semantic role features for machine translation // Proceedings of the 23rd International Conference on Computational Linguistics. — Association for Computational Linguistics, 2010. — P. 716–724.
9. Relation alignment for textual entailment recognition / Mark Sammons, VG Vinod Vydiswaran, Tim Vieira et al. // Text Analysis Conference (TAC). — 2009.
10. *Xue N., Palmer M.* Calibrating features for semantic role labeling // Proceedings of EMNLP 2004. — Association for Computational Linguistics, 2004. — P. 88–94.
11. Shallow semantic parsing using support vector machines / Sameer S Pradhan, Wayne H Ward, Kadri Hacioglu et al. // HLT-NAACL 2004: Main Proceedings. — Association for Computational Linguistics, 2004. — P. 233–240.
12. *Toutanova K., Haghighi A., Manning C. D.* Joint learning improves semantic role labeling // Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. — Association for Computational Linguistics, 2005. — P. 589–596.
13. *Punyakanok V., Roth D., Yih W.-t.* The importance of syntactic parsing and inference in semantic role labeling // Computational Linguistics. — 2008. — Vol. 34, no. 2. — P. 257–287.
14. *Palmer M., Gildea D., Kingsbury P.* The proposition bank: An annotated corpus of semantic roles // Computational linguistics. — 2005. — Vol. 31, no. 1. — P. 71–106.
15. *Fillmore C. J., Johnson C. R., Petruck M. R.* Background to FrameNet // International journal of lexicography. — 2003. — Vol. 16, no. 3. — P. 235–250.
16. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages / Jan Hajic, Massimiliano Ciaramita, Richard Johansson et al. // Proceedings of the Thirteenth Conference on Computational Natural

Language Learning: Shared Task. — Association for Computational Linguistics, 2009. — P. 1–18.
17. *Fung P., Chen B.* BiFrameNet: bilingual frame semantics resource construction by cross-lingual induction // Proceedings of the 20th international conference on Computational Linguistics. — Association for Computational Linguistics, 2004.
18. Cross-language frame semantics transfer in bilingual corpora / Roberto Basili, Diego De Cao, Danilo Croce et al. // International Conference on Intelligent Text Processing and Computational Linguistics / Springer. — 2009. — P. 332–345.
19. *Padó S., Lapata M.* Cross-lingual annotation projection for semantic roles // Journal of Artificial Intelligence Research. — 2009. — Vol. 36. — P. 307–340.
20. *Johansson R., Nugues P. A* FrameNet-based semantic role labeler for Swedish // Proceedings of the COLING/ACL. — Association for Computational Linguistics, 2006. — P. 436–443.
21. *Kozhevnikov M., Titov I.* Cross-lingual transfer of semantic role labeling models // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Association for Computational Linguistics, 2013. — P. 1190–1200.
22. *Das D., Smith N. A.* Semi-supervised frame-semantic parsing for unknown predicates // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. — Association for Computational Linguistics, 2011. — P. 1435–1444.
23. *Burchardt A., Erk K., Frank A.* A WordNet detour to FrameNet // Sprachtechnologie, mobile Kommunikation und linguistische Resourcen. — 2005. — Vol. 8. — P. 408–421.
24. *Miller G. A.* WordNet: A lexical database for English // Communications of the ACM. — 1995. — Vol. 38, no. 1. — P. 39–41.
25. *Johansson R., Nugues P.* Using WordNet to extend FrameNet coverage // In Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages at the 16th Nordic Conference of Computational Linguistics (NODALIDA). — 2007. — P. 27–30.
26. Automatic induction of FrameNet lexical units / Marco Pennacchiotti, Diego De Cao, Roberto Basili et al. // Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. — Association for Computational Linguistics, 2008.
27. *Fürstenau H., Lapata M.* Semi-supervised semantic role labeling via structural alignment // Computational Linguistics. — 2012. — Vol. 38, no. 1. — P. 135–171.
28. *Do Q. T. N., Bethard S., Moens M.-F.* Domain adaptation in semantic role labeling using a neural language model and linguistic resources // IEEE/ACM Transactions on Audio, Speech, and Language Processing. — 2015. — Vol. 23, no. 11. — P. 1812–1823.
29. *Garg N., Henderson J.* Unsupervised semantic role induction with global role ordering // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. — Association for Computational Linguistics, 2012. — P. 145–149.
30. *Lang J., Lapata M.* Similarity-driven semantic role induction via graph partitioning // Computational linguistics. — 2014. — Vol. 40, no. 3. — P. 633–669.
31. *Titov I., Khoddam E.* Unsupervised induction of semantic roles within a reconstruction error minimization framework // In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — 2015.
32. *Shelmanov A. O., Smirnov I. V.* Methods for semantic role labeling of Russian texts // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (2014). — No. 13. — 2014. — P. 607–620.
33. *Kuznetsov I.* Semantic role labeling for Russian language based on Russian FrameBank // International Conference on Analysis of Images, Social Networks and Texts / Springer. — 2015. — P. 333–338.
34. *Sokirko, A. V.* 2001. Semanticheskie slovari v avtomaticheskoy obrabotke teksta (po materialam sistemy DI-ALING) [Semantic dictionaries in automatic text processing]. PhD Thesis. Moscow.
35. *Osipov, G. S., Shelmanov, A. O.* 2015. Metod povysheniya kachestva sintaksicheskogo analiza na osnove vzaimodeystviya sintaksicheskikh i semanticheskikh pravil [Method of improving the quality of parsing based on the interaction of syntactic and semantic rules]. Trudy shestoy mezhdunarodnoy konfe-rentsii “Sistemnyy analiz i informatsionnye tekhnologii” (SAIT) [6th Conference “Systems Analysis and Information Technologies”]. Svetlogorsk. p. 229–240.
36. *Smirnov, I. V., Shelmanov, A. O., Kuznetsova, E. S., Khramoin, I. V.* Semantiko-sintaksicheskiy

analiz estestvennykh yazykov. Chast' II. Metod semantiko-sintaksicheskogo analiza tekstov [The semantic-syntactic analysis of natural languages. Part II. The method of semantic and syntactic analysis of texts]. Is-kusstvennyy intellekt i prinyatie resheniy [Artificial intelligence and decision-making]. 1: 11–24.

37. *Osipov G. S., Smirnov I. V., Tikhomirov I. A.* Reliacionno-situacionnyi metod poiska i Analisa tekstov I ego prilogienia // Iskusstvennyy intellekt i prinyatie resheniy [Artificial intelligence and decision-making]. 2008. — No 2. — p. 3–10.
38. *Zolotova, G.A., Onipenko, N.K., Sidorova, M.Yu.* 2004. Kommunikativnaya grammatika russkogo yazyka [Communicative Grammar of the Russian Language] // Moscow: Russian Vinogradov Language Institute of RAS. 544 p.
39. *Apresyan, Yu. D., Boguslavskiy, I. M., Iomdin, B. L.*, i dr. 2005. Sintaksicheski i semanticheski annotirovannyy korpus russkogo yazyka: sovremennoe sostoyanie i perspektivy [Syntactically and semantically annotated corpus of Russian language: current status and prospects]. Natsional'nyy korpus rus-skogo yazyka [National Corpus of Russian Language]. P. 193–214.
40. Avtomaticheskaya obrabotka teksta [Automatic Text Processing]. Available at: http://www.aot.ru/ (Accesssed November 20, 2016).
41. MaltParser: A language-independent system for data-driven dependency parsing / Joakim Nivre, Johan Hall, Jens Nilsson et al. // Natural Language Engineering. — 2007. — Vol. 13, no. 2. — P. 95–135.
42. Distributed representations of words and phrases and their compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // Advances in neural information processing systems. — 2013. — P. 3111–3119.
43. *Mnih A., Kavukcuoglu K.* Learning word embeddings efficiently with noise-contrastive estimation // Advances in Neural Information Processing Systems. — 2013. — P. 2265–2273.
44. *Kutuzov A., Andreev I.* Texts in, meaning out: neural language models in semantic similarity task for Russian // Proceedings of the Dialog Conference. — 2015.

**Shelmanov Artem Olegovich**, Fellow of ISA FRC CSC RAS. PhD Graduated from National Research Nuclear University «MEPhI» in 2011. The number of publications: 20. Research interests: artificial intelligence, natural language processing, machine learning, search and analytical systems. E-mail: shelmanov@isa.ru

**Kamenskaya Margarita Alexandrovna**, research engineer of ISA FRC CSC RAS. Graduated from Peoples' Friendship University of Russia in 2014. The number of publications: 5. Research interests: natural language processing, computational linguistics, coreference resolution. E-mail: mak@isa.ru