На правах рукописи

Кузнецов Илья Олегович

Автоматическая разметка семантических ролей в русском языке

10.02.21 — Прикладная и математическая лингвистика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата филологических наук

Москва - 2016

Работа выполнена в Школе лингвистики факультета гуманитарных наук Научно-исследовательского университета Высшая школа экономики.

Научный руководитель: кандидат филологических наук
доцент Школы лингвистики факультета гуманитарных наук НИУ ВШЭ
Бонч-Осмоловская Анастасия Александровна

Официальные оппоненты: **Соловьев Валерий Дмитриевич**
доктор физико-математических наук,
профессор Казанского федерального университета

Иомдин Леонид Лейбович
кандидат филологических наук,
ведущий научный сотрудник Лаборатории №15
Института проблем передачи информации РАН

Ведущая организация: Институт проблем информатики
Российской академии наук
Федерального исследовательского центра
"Информатика и управление" Российской академии наук

Защита состоится 11 мая 2016 года в 16:00 часов на заседании диссертационного совета Д 501.001.24 при Московском государственном университете имени М.В. Ломоносова по адресу: 119991, г. Москва, ГСП-1, Ленинские горы, МГУ имени М.В. Ломоносова, 1-й учебный корпус, филологический факультет.
С диссертацией можно ознакомиться в научной библиотеке Московского государственного университета имени М. В. Ломоносова.

Автореферат разослан 9 марта 2016 года.

Учёный секретарь
диссертационного совета
доктор филологических наук А. М. Белов

Общая характеристика исследования

Объект предложенного диссертационного исследования - автоматическая разметка актантов методами машинного обучения для русского языка. Автоматическая разметка семантических ролей, или **автоматическая разметка актантов** (*Semantic Role Labeling, SRL*) - одно из приоритетных направлений в современной автоматической обработке языка. Это тип высокоуровневого анализа текста, при котором для исходного текста на естественном языке порождается поверхностная интерпретация на основе теории семантических ролей.

Предположим, что дано предложение на естественном языке, и в этом предложении выбран некоторый **предикат** (например, глагол). Задача автоматической разметки актантов состоит в том, чтобы найти в предложении именные группы, обозначающие участников описанной предикатом ситуаци (актанты) и приписать им **семантические роли**. Так, например, предложение "*Пётр купил яблоко за 5 рублей*" будет проанализировано следующим образом:

$$[\text{Пётр}]_{\text{Покупатель}}\ \text{купил}\ [\text{яблоко}]_{\text{Товар}}\ \text{за}\ [\text{5 рублей}]_{\text{Цена}}$$

Автоматическая разметка актантов отличается от синтаксического парсинга, в ходе которого анализу подвергается грамматика, а не семантика высказывания, и от полного семантического анализа, т.к. работа всегда производится на уровне предложения, и системы не используют правил логического вывода. Анализ текста в терминах семантических ролей позволяет сравнительно небольшими усилиями получить дополнительный уровень абстракции, описывающий семантику текста. Информация о семантических ролях может быть затем использована для извлечения фактов, для машинного перевода, в вопросно-ответных системах, а также, потенциально, в любой системе автоматической обработки языка, которая так или иначе опирается на семантическую информацию.

Автоматическая разметка актантов в современном понимании возникла в начале 2000-х годов и была описана в работах Д. Журафски и Д. Гилдеа[1]. Теоретической основой для направления послужила **теория семантических ролей** Ч. Филлмора[2]. Прикладным основанием экспериментов в этой области можно считать построенные на базе теории Филлмора лексико-грамматические ресурсы: в первую очередь, это модели FrameNet[3], PropBank[4] и VerbNet[5]. Традиционно автоматическая обработка актантов опирается на ряд синтаксических, морфологических и лексических признаков для принятия решения о том, какую роль следует приписать выбранному участнику ситуации. Так, например, в работе Д. Гилдеа и Д. Журафски для этого использовался путь в дереве составляющих от предиката до выбранной именной группы, залог глагола, кластер лексемы, выражающей участника, и др. Для того чтобы получить доступ к этим свойствам, требуется предварительно произвести морфологический анализ текста, лемматизацию, синтаксический анализ и т.д. Создание систем, которые выполняли бы подобный анализ, – отдельная и сложная задача, и для большинства языков подобные системы отсутствуют. Кроме того, для автоматической разметки актантов методами машинного обучения требуется создать обучающий корпус примеров, размеченных по семантическим ролям. Создание такого корпуса – также крайне трудоёмкая задача.

В последние годы было проведено множество исследований по автоматической обработке текстов для русского языка. Однако, несмотря на общую популярность, тема автоматической разметки актантов почти не исследовалась на русском материале, и одной из причин этого было отсутствие обучающего и тестового корпуса с разметкой по семантическим ролям и доступных инструментов предварительной обработки текста. На

1 Gildea, D., Jurafsky, D. (2000). Automatic labeling of semantic roles. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics - ACL ’00, (1972), 512–520

2 Fillmore, C. J. (1968). The Case for Case. In E. Bach & R. T. Harms (Eds.), Universals in Linguistic Theory (pp. 0–88). New York: Holt, Rinehart and Winston.

3 C. F. Baker, Fillmore, C. J., and Lowe, J. B., “The Berkeley FrameNet project”, in COLING-ACL '98: Proceedings of the Conference, Montreal, Canada, 1998, pp. 86-90.

4 Palmer M, Kingsbury P, Gildea D (2005). "The Proposition Bank: An Annotated Corpus of Semantic Roles". Computational Linguistics 31 (1): 71–106.

5 Schuler, K. K. (2005). VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. Dissertation Abstracts International, B: Sciences and Engineering, 66(6).

сегодняшний день активно разрабатывается ресурс FrameBank[6], один из компонентов которого представляет собой корпус с необходимой для нашей задачи разметкой. В рамках диссертационного исследования мы разработали систему автоматической разметки актантов, опираясь на промежуточную версию этого ресурса. Подобной работы на материале FrameBank ранее не проводилось.

Цель представленного исследования – разработать и описать систему автоматической разметки актантов и детально изучить результаты её работы, выяснить вклад различных лингвистических свойств и других параметров задачи в качество классификации. В рамках исследования мы выделяем следующие подзадачи:

- Интегрировать доступные ресурсы предобработки в цепочку, которая позволит обогатить исходный корпус FrameBank морфологической и синтаксической информацией
- Произвести фильтрацию корпуса примеров FrameBank, обеспечив тем самым высокое качество обучающих и тестовых данных
- Разработать модель для классификации актантов на основе деревьев зависимостей и лингвистических свойств, в т.ч. специфичных для русского языка
- Разработать модуль глобальной оптимизации, который обеспечивает выполнение ограничений, накладываемых теорией семантических ролей
- Оценить качество работы полученной системы на изолированной тестовой выборке. Оценить вклад лингвистических свойств и других параметров задачи в качество работы системы.
- Выработать рекомендации по дальнейшему развитию системы и корпуса FrameBank.

В качестве **материала** исследования мы используем корпус примеров FrameBank, а также на построенные на основе этого корпуса модели. Автоматическая разметка актантов для русского языка – одно из наименее развитых направлений в

6 Lyashevskaya, O., Kashkin, E. (2015). FrameBank: A Database of Russian Lexical Constructions for the "Deep" Parsing of Russian. Analysis of Images, Social Networks and Texts. 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers, 542.

автоматической обработке текста, что, учитывая большое прикладное значение этой задачи, объясняет её **актуальность**. **Научная новизна** работы состоит в том, что ранее подобных исследований на русском материале не проводилось. Предложенное исследование – первый опыт применения систем на основе машинного обучения к корпусу примеров FrameBank. Ряд частных решений также применяется к русскому языку впервые, кроме того, это первое известное нам полноценное описание подобной системы, достаточно подробное для успешной реимплементации и усовершенствования предложенного метода.

Теоретическая значимость исследования состоит в оценке вклада различных лингвистических свойств в качество работы классификатора. Система опирается на ряд свойств, в т.ч. на путь в дереве зависимостей, падеж актанта, предлог, которым оформлен актант, лемму и кластер актанта и др. Наш анализ демонстрирует важность синтаксических свойств для автоматической разметки актантов в русском языке. Роль лексических свойств оказывается второстепенной, и мы подробно рассматриваем возможные причины такого поведения системы. Для русского языка подобное исследование проводится впервые.

Практическая значимость исследования состоит в подробном качественном и количественном анализе результатов работы системы. Кроме того, работа содержит детальное описание компонентов системы, а также ряд рекомендаций для дальнейших экспериментальных исследований в этой области.

Апробация работы. Основные положения исследования и полученные результаты были представлены на конференциях «Анализ Изображений, Сетей и Текстов» (АИСТ-2013), Analysis of Images, Social Networks and Texts (AIST-2015) на Международной конференции «Диалог» (2013). Также по материалам исследования было опубликовано три статьи в журнале «Научно-техническая информация» (2012 и 2013 гг.).

Структура диссертации. Диссертация состоит из введения, четырёх глав, заключения и библиографии. **Глава I** посвящена теоретическим основам автоматической разметки актантов, а также истории этого направления. **Глава II** содержит описание разработанной системы автоматической разметки актантов. **Глава**

III посвящена процедуре оценки качества работы системы и описанию полученных результатов. **Глава IV** подводит итоги проведённого исследования. Заключение завершает работу.

Основное содержание работы

Во **Введении** приводится общее описание исследовательской задачи, указываются основные методы решения задачи и возникающие при этом сложности. Также даётся обоснование актуальности выбранной темы, её научной новизны, теоретической и практической значимости.

Глава I

Теория семантических ролей и автоматическая разметка актантов

В теоретическом отношении автоматическая разметка актантов опирается на теорию семантических ролей. Глава I посвящена истории и современному состоянию теории семантических ролей и автоматической разметки актантов.

Понятие семантической роли, которое используется в современной автоматической обработке актантов, основывается на работах Ч. Филлмора, который ввёл понятие семантической роли в современный лингвистический дискурс, и Дж. Грубера, который оперировал концептуально схожим понятием тематического отношения. Классическая теория семантических ролей, предложенная Филлмором, постулирует наличие инвентаря семантических ролей, обладающих следующими свойствами:

- **Полнота** и **уникальность** - *каждый* аргумент глагола имеет *ровно одну* семантическую роль
- **Единственность заполнения** - роль может быть заполнена только один раз
- **Независимость** и **атомарность** - семантическая роль имеет категориальную природу и не может быть разделена на компоненты.

Классический инвентарь семантических ролей включает в себя такие роли как *Агенс*, *Пациенс*, *Бенефактив*, *Инструмент*, *Экспериенцер*, *Стимул* и др. В ходе дальнейших исследований семантических ролей выяснилось, однако, что этот инвентарь обладает ограниченными описательными возможностями и что ни одно из указанных выше свойств не является абсолютным.

На сегодняшний день существует три основных подхода к созданию инвентаря семантических ролей. Первый подход использует наиболее дробное представление ролей, в котором роли являются **предикатно-специфическими**, т.е. уникальными для каждого предиката: например, у глагола "убивать" будут представлены роли "тот, кто убивает", "тот, кого убивают", "орудие убийства" и т.д. На другом конце спектра находятся подходы, опирающиеся на максимально **обобщённые роли** Актора и Претерпевающего: эти роли отвечают за большую долю вариативности в синтаксическом поведении аргументов, и использование крупных ролей открывает возможности для генерализации, недоступные для более "дробных" инвентарей, в то же время понижая внутреннюю семантическую однородность ролей. Наконец, в середине спектра находятся **классические** ролевые инвентари наподобие предложенного Ч. Филлмором.

В контексте автоматической разметки актантов наибольшую популярность имеют подходы на основе предикатно-специфических ролей.

При экспертной разметке корпуса примеров FrameBank, послужившего материалом данного диссертационного исследования, используется формализм, разработанный в рамках Московской семантической школы. Исходя из того, что понятие актанта в МСШ и понятие предикатно-специфичной семантической роли функционально близки, мы ставим перед собой задачу автоматической разметки актантов - или автоматической разметки семантических ролей, и в дальнейшем используем два этих понятия как взаимозаменяемые, хотя с теоретической точки зрения это не совсем соответствует действительности. При разработке системы автоматческой разметки актантов мы опираемся на характеристики семантических ролей, которые традиционно используются в Semantic Role Labeling, и моделируем синтаксическое

оформление актантов, ограничения на лексическое заполнение валентностей, устойчивость к трансформациям и ограничение на единственность заполнения роли.

Глава содержит исторический обзор систем автоматической разметки семантических ролей для английского языка.

Параллельно с уже упомянутой выше работой Д. Гилдеа и Д. Журафски, посвящённой автоматической разметке актантов с использованием ролей FrameNet, увидела свет работа Д. Гилдеа и М. Палмер[7], посвящённая разметке семантических ролей на основе корпуса PropBank с сопоставимыми результатами. В 2004 и 2005 годах в рамках конференции CoNLL были проведены соревнования по автоматической разметке актантов. В рамках соревнований автоматическая разметка актантов производилась на материале английского языка с использованием синтаксиса непосредственных составляющих. В качестве исходных данных системам был предложен корпус PropBank.

Лучший результат на соревновании CoNLL-2005 продемонстрировала система В. Пуньяканок[8]. Архитектура этой системы состоит из трёх модулей: идентификации актантов, присвоения ролей и дополнительного модуля глобальной оптимизации на основе целочисленного программирования. Другая интересная работа, также представленная в рамках CoNLL-2005, - исследование М. Сурдеану и Дж. Турмо[9], посвящённое сравнению качества работы систем SRL на основе полного и частичного синтаксического разбора. Эта работа продемонстрировала, что, несмотря на ошибки синтаксического анализатора, использование полного синтаксического анализа позволяет получить лучшие или по крайней мере сопоставимые результаты. Работа С. Прадхан и др.[10] демонстрирует альтернативный подход, в котором Semantic Role Labeling интерпретируется как задача сегментации.

---

7 Gildea D., Palmer M. The necessity of parsing for predicate argument recognition // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics — 2002. — № July — С. 239–246.

8 Koomen, P., Punyakanok, V., Roth, D., & Yih, W. (2005). Generalized inference with multiple semantic role labeling systems. Proceedings of the Ninth Conference on Computational Natural Language Learning, 181–184.

9 Surdeanu, M., & Turmo, J. (2005). Semantic role labeling using complete syntactic analysis // CONLL '05 Proceedings of the Ninth Conference on Computational Natural Language Learning

10 Pradhan, S., Hacioglu, K., Ward, W., Martin, J. H., & Jurafsky, D. (2005). Semantic role chunking combining complementary syntactic views. Proceedings of the Ninth Conference on Computational Natural Language Learning - CONLL ’05,

В последующий период были предприняты попытки как улучшить существующие результаты для английского языка, так и разработать системы автоматической классификации актантов для других языков. В ходе этих исследований выяснилось, что синтаксис непосредственных составляющих недостаточно удобен для представления синтаксической информации в языках со свободным порядком слов и падежным маркированием. Было продемонстрировано, что синтаксис деревьев зависимостей в таких случаях обладает большей описательной силой.

Кроме того, было показано, что связь между задачами синтаксического и поверхностного семантического анализа — двусторонняя: не только автоматическая разметка актантов опирается на синтаксис, но и наоборот, синтаксический анализ может быть выполнен с лучшим качеством, если предоставить системе данные о семантических ролях. Один из первых подходов, в котором синтаксический и семантический анализ оказываются взаимозависимы, был предложен уже в 2005 году в работе К. Тутановой и К. Маннинга[11]. Авторы использовали классификатор на основе максимальной энтропии со стандартным набором свойств на основе деревьев непосредственных составляющих, однако вместо единственного синтаксического представления классификация актантов выполнялась на ранжированном наборе синтаксических разборов, полученных автоматически.

Указанные выше тенденции привели к появлению нового типа систем, которые основывались на синтаксисе деревьев зависимостей. В 2007 и 2008 году были проведены соревнования CoNLL 2007 и 2008, посвящённые задаче автоматического синтаксического и семантического анализа как для английского, так и других языков.

Современные системы автоматической классификации актантов опираются на более сложные методы, в которых информация о структуре задачи и особенностях семантического представления кодируется непосредственно в модели. В качестве примера такой системы можно привести систему SEMAFOR[12]. Все рассмотренные нами ранее системы представляют собой последовательность независимых классификаторов.

11 Haghighi, A., Toutanova, K., & Manning, C. (2005). A joint model for semantic role labeling. Proceedings of the Ninth Conference on Computational Natural Language Learning - CONLL '05,

12 Das, D. (2014). Statistical Models for Frame-Semantic Parsing. Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014), (2007), 26–29

Один из недостатков такого подхода — невозможность использовать информацию о присвоении ролей на этапе идентификации актантов. В системе SEMAFOR идентификация и классификация актантов выполняются одновременно.

На сегодняшний день практически не имеется публикаций, посвященных решению задачи автоматической классификации актантов на русском материале. Единственная известная нам реализация данной задачи на основе машинного обучения И. Смирнова и А. Шелманова[13] относится к методам частичного обучения с учителем и не использует корпус с разметкой по семантическим ролям. Также для русского языка существует несколько систем на основе правил[14] и систем извлечения фактов, в том числе позволяющих генерировать шаблоны описания событий на основе больших массивов неразмеченных данных[15].

Глава II

Система автоматической разметки актантов для русского языка

В Главе II приводится подробное описание системы автоматической разметки актантов для русского языка, разработанной в ходе диссертационного исследования. Подробно рассматриваются использованные в системе методы машинного обучения, лингвистические свойства, на основе которых происходит классификация, а также ряд технических решений, использованных при создании системы и работе с исходными данными. Глава включает в себя несколько разделов.

Раздел 1 посвящен обзору различных подходов к решению задачи автоматической разметки актантов, выбору оптимального подхода и обоснованию этого выбора.

13 Смирнов, И. В., Шелманов, А. О. (2014). Методы установления семантических ролей для текстов на русском языке // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4 — 8 июня 2014 г.). — Москва: РГГУ, 2014. — С. 607–619.

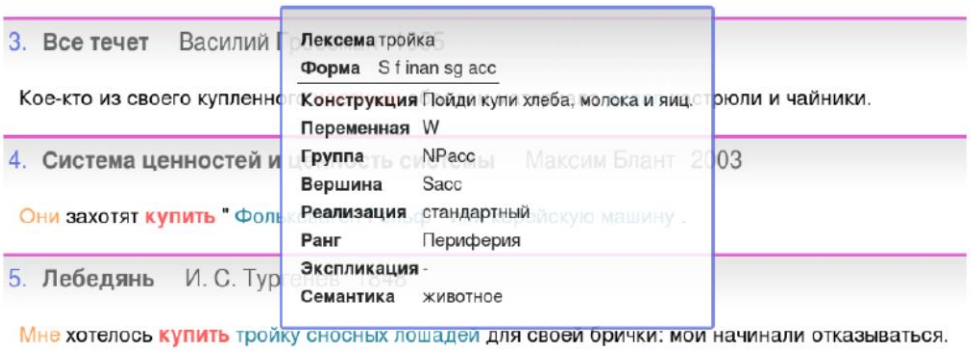
14 Анисимович, К. В., Дружкин, К. Ю., Зуев, К. А., & Петрова, М. А. (2012). Синтаксический И Семантический Парсер, Основанный На Лингвистических Технологиях Abbyy Compreno. Международная конференция по компьютерной лингвистике «Диалог-2012»

15 Котельников, Д. С., & Лукашевич, Н. В. (2012). Итерационное извлечение шаблонов описания событий по новостным кластерам. Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

Предложенная в диссертационном исследовании система основана на **предикатно-специфических ролях**, т.к. именно этот тип ролей использовался при разметке корпуса примеров FrameBank, послужившего материалом для диссертационного исследования. В предложенной нами системе задача актантов трактуется как задача **классификации узлов деревьев зависимостей**, т.к. данный подход является более естественным для языков со свободным порядком слов, каким является русский язык. Предложенная система также содержит **модуль глобальной оптимизации** на основе целочисленного программирования. Решение о том, является ли узел актантом, и о том, какую семантическую роль он имеет, принимается совместно, в один шаг, а не в два, как это делалось в ранних системах автоматической разметки семантических ролей.

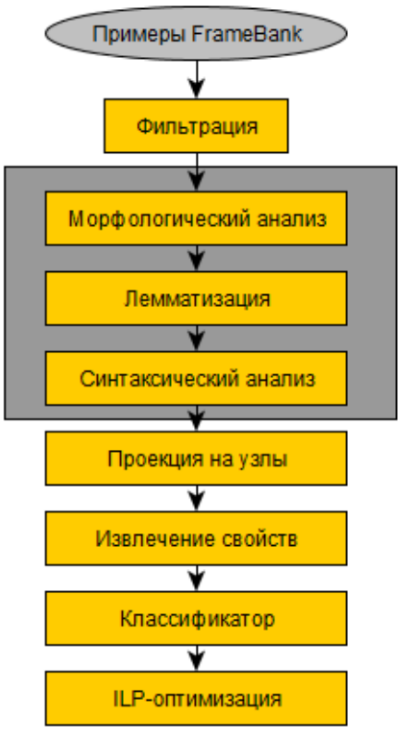
Раздел 2 Главы II описывает **исходные данные**, которые были использованы для обучения и тестирования классификатора: коллекцию примеров из корпуса FrameBank. FrameBank представляет собой корпусно-лексикографический ресурс, описывающий лексические конструкции русского языка с помощью специальным образом размеченных предложений из Национального корпуса русского языка. Центральным организующим компонентом системы разметки, используемой в FrameBank, является **лексическая конструкция.**

Каждой конструкции в системе FrameBank соответствует набор примеров из Национального корпуса русского языка (НКРЯ). Примеры представляют собой отрывки текста, разбитые на предложения и слова. Для каждого слова дана морфологическая информация, полученная с помощью автоматического анализатора, а также семантические пометы из инвентаря НКРЯ. Разметка по семантическим ролям производится в первую очередь для глагольных конструкций, включая конструкции с нефинитными формами (причастиями, деепричастиями, инфинитивами и т.д.), что увеличивает сложность распознавания семантических ролей в рамках одной конструкции. Следующий пример демонстрирует разметку, которая используется в качестве исходных данных для предложенной в работе системы.



Эта разметка поступает на вход модуля извлечения свойств, который преобразует разметку в признаки для классификации.

Раздел 3 посвящён описанию разработанной системы. Архитектура системы включает в себя следующие модули: модуль препроцессинга (фильтрация, морфологический анализ, лемматизация, синтаксический анализ), модуль обогащения данных (проекция разметки на узлы дерева зависимостей), модуль обучения (извлечение свойств, классификатор, ILP-оптимизация). Приведённая ниже схема иллюстрирует взаимодействие модулей системы.



На вход системы поступает база данных FrameBank, которая помимо прочего содержит размеченные по семантическим ролям примеры употребления конструкций из

Национального Корпуса русского языка в формате xml. Поскольку ресурс находится на стадии разработки, некоторые примеры в корпусе содержат ошибки разметки, связанные в большинстве случаев с техническими причинами. Для того чтобы дальнейшая работа была возможной, мы применяем процедуру фильтрации корпуса, в результате которой на основании простых правил принимаем решение, какие из предложений будут использованы в эксперименте.

Предложения корпуса FrameBank разбиты на слова и содержат слой морфологической разметки. Для автоматической разметки актантов, однако, важную роль играет синтаксический анализ текста, поэтому мы дополнительно обрабатываем исходные данные **синтаксическим парсером** и обеспечиваем интеграцию исходного корпуса и полученных синтаксических деревьев.

Учитывая, что в предложенной системе задача автоматической классификации актантов интерпретируется как задача разметки узлов, необходимо осуществить отображение разметки FrameBank с отрезков текста на узлы соответствующего синтаксического дерева. За эту операцию отвечает **модуль проекции на узлы.**

Далее все предложения-примеры из корпуса группируются по конструкциям, которые они описывают, формируя таким образом подкорпуса примеров для каждой отдельной конструкции.

Для каждого из полученных подкорпусов производится случайное разбиение на тренировочную и тестовую выборки. Единицей разбиения мы принимаем предложение. Тренировочная выборка поступает на вход классификатора и используется для обучения, тестовая выборка используется для оценки качества работы классификатора.

Как тренировочная, так и тестовая выборка поступают на вход модуля извлечения свойств, который преобразует информацию, полученную в результате предварительной обработки, в свойства, используемые классификатором. Модуль **извлечения свойств** приписывает набор признаков каждому узлу дерева зависимостей, построенного для каждого предложения тренировочной и тестовой выборки. Здесь же экземпляры-узлы получают метку класса: в тренировочной выборке эта метка используется для обучения классификатора, а в тестовой – для сравнения результатов работы системы с эталонной разметкой.

На этапе тестирования каждое предложение тестовой выборки подаётся на вход **классификатору**, который для каждого узла в дереве зависимостей этого предложения определяет его семантическую роль. В результате такой независимой классификации одна и та же роль может быть приписана нескольким узлам, что противоречит базовым принципам теории семантических ролей и нежелательно с практической точки зрения. Для решения этой проблемы используется модуль **ILP-оптимизации** на основе метода целочисленного программирования для постобработки результатов классификации. Результат работы модуля оптимизации является конечным результатом работы системы и поступает на выход.

Отдельное внимание в Разделе 3 уделено свойствам, использованным для обучения классификатора. В работе используются свойства, традиционно применяемые в системах автоматической классификации актантов на основе машинного обучения. Каждый узел дерева зависимостей представлен в терминах следующих свойств:

- **синтаксические свойства**
  - путь (*path*) – путь от предиката до актанта в дереве зависимостей
  - короткий путь (*shortPath4*) – сокращённый путь от предиката до актанта
  - падеж (*case*) – падеж актанта
  - финский падеж (*finncase*) – падеж + предлог актанта
  - форма глагола (*vform*)
  - залог (*voice*)
- **семантические свойства**
  - лемма (*lemma*)
  - кластер (*cluster*) – кластер, к которому принадлежит лемма актанта. Кластеры получены с помощью алгоритма Chinese Whispers, применённого к векторной модели на основе word2vec из проекта RusVectores[16]
  - часть речи (*POS*)

16 Kutuzov, A., & Andreev, I. (2015). Texts in, Meaning Out: Neural Language Models in Semantic Similarity Tasks for Russian // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог» (Москва, 27 — 30 Мая 2015 г.).

Раздел 4 описывает использованный в работе модуль глобальной оптимизации на основе целочисленного программирования. Данный модуль оптимизирует результаты классификации узлов таким образом, чтобы максимизировать уверенность классификатора при соблюдении определённых лингвистически мотивированных ограничений. Раздел включает в себя общее описание технологии, а также детали реализации модуля. Как упоминалось выше, каждая семантическая роль может быть приписана только одному узлу в предложении. Дополнительно требуется, чтобы такое приписание обеспечивало максимальную «уверенность» системы в предоставленных ответах.

Соответствующая задача целочисленного программирования формулируется следующим образом. Пусть у нас имеется набор ролей $\{r_1, r_2, r_3 \dots r_i\}$ и набор узлов $\{n_1, n_2, n_3 \dots n_j\}$. Одна из ролей является "пустой" ролью и обозначает отсутствие семантической роли на выбранном узле. Эта роль имеет особый статус, т.к., в отличие от других ролей, может быть заполнена неограниченное число раз. Обозначим её как $r_\emptyset$. Введём набор переменных-индикаторов x_{ij} , которые обозначают, что роль r_i приписана узлу n_j, и набор переменных w_{ij}, в которых хранится вес роли r_i для узла n_j , определённый классификатором на основе свойств, выделенных для данного узла. Следующий пример иллюстрирует семантику переменной:

Иван купил яблоко

	r_1(Покупатель)	r_2(Товар)	$r_\emptyset$
n_1(Иван)	$w_{11} = 0.8$	$w_{21} = 0.2$	$w_{\emptyset 1} = 0.1$
n_2(Яблоко)	$w_{12} = 0.4$	$w_{22} = 0.7$	$w_{\emptyset 2} = 0.2$

Задача оптимизации состоит в следующем. Необходимо выбрать значения переменных x_{ij} таким образом, чтобы максимизировать функцию уверенности $\sum_{i,j} x_{ij} w_{ij}$. При этом существует два принципиальных ограничения, которые должны быть соблюдены. Во-первых, одному узлу может быть приписана только одна роль. Для каждого узла n_j необходимо обеспечить $\forall j$: $\sum_i x_{ij} = 1$. За счёт того, что переменные

x_{ij} целые и принимают значения $\{0,1\}$, это условие может быть верно только когда узел имеет единственную роль. Во-вторых, каждая роль, за исключением "пустой" роли, может быть приписана только один раз. Это условие выражается в терминах наших переменных следующим образом: $\forall i \neq \emptyset$: $\sum_j x_{ij} \leq 1$.

Задача оптимизации передаётся в модуль линейного программирования, который решает её с помощью симплексного метода. Общая формулировка задачи для случая с двумя переменными приводится ниже.

Максимизировать:
$x_{11}w_{11} + x_{12}w_{12} + x_{21}w_{21} + x_{22}w_{22} + x_{\emptyset 1}w_{\emptyset 1} + x_{\emptyset 2}w_{\emptyset 2}$
С учётом ограничений:
“каждый узел получает одну роль”
$x_{11} + x_{21} + x_{\emptyset 1} = 1$
$x_{12} + x_{22} + x_{\emptyset 2} = 1$
“каждая роль заполняется максимум один раз”
$x_{11} + x_{12} \leq 1$
$x_{21} + x_{22} \leq 1$

Сформулированная таким образом задача поступает в модуль целочисленного программирования, который выбирает значения переменных, удовлетворяющие указанным требованиям.

Раздел 5 Главы II посвящён техническим особенностям имплементации системы. Для русского языка не существует общепринятого единого набора инструментов предварительной обработки текста, и значительная часть усилий в ходе подготовки к исследованию была затрачена на поиск и интеграцию различных компонентов автоматической обработки текста. Раздел перечисляет использованные компоненты и содержит множество технических деталей и рекомендаций, которые могут упростить разработку аналогичных систем в будущем.

Глава III

Экспериментальная оценка и результаты

Глава III посвящена экспериментальной оценке качества разработанной системы. Для задач машинного обучения эта область хорошо разработана и существует ряд стандартных параметров, по которым можно определить, насколько хорошо работает

система. В рамках диссертационного исследования оценка качества выполнялась на основании тестовой выборки. В наиболее интересных случаях был произведён экспертный анализ результатов.

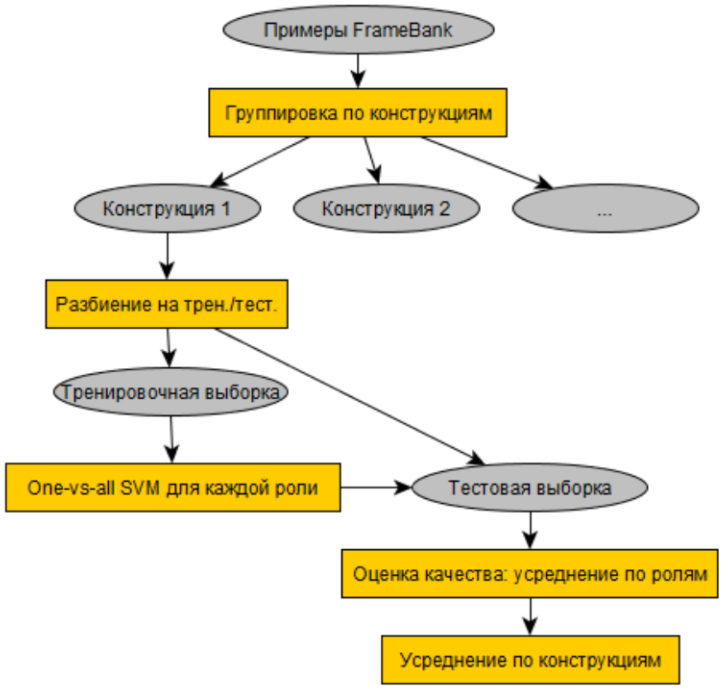
При количественной оценке работы системы в работе используются меры **точность** (Precision, P), **полнота** (Recall, R) и **F-мера.**

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F_1 = \frac{2PR}{P + R}$$

Эти показатели вычисляются отдельно для каждого класса, и затем полученные показатели усредняются по всем классам.

В ходе оценки предложения исходного корпуса были сгруппированы в зависимости от того, какую конструкцию они описывают. Каждый из полученных подкорпусов в свою очередь случайным образом разбивается на тестовую и тренировочную выборки. Затем, с использованием тренировочной выборки производится обучение классификаторов типа "один против всех" для каждой роли. При использовании системы на тестовых данных каждый экземпляр передаётся каждому из бинарных классификаторов, которые, в свою очередь, возвращают вес - меру “уверенности” классификатра в том, что данный экземпляр принадлежит к его классу. Затем система выбирает класс с наибольшим весом и приписывает его экземпляру. В случае с ILP-постобработкой выбор классов производится путём решения LP-задачи оптимизации и максимизирует суммарную "уверенность" классификатора на всём предложении. В любом случае, в результате применения системы каждый экземпляр (узел дерева зависимостей) входного предложения получает одну из ролевых меток. На основании этих данных для каждой роли рассчитывается точность, полнота и F-мера. Эти меры усредняются по всем ролям, и полученный результат считается результатом работы системы для выбранной конструкции. Затем значения мер усредняются ещё раз по всем конструкциям, и полученные средние для точности, полноты и F-меры в дальнейшем считаются "качеством работы системы в данной конфигурации".

Следующая схема демонстрирует структуру использованной процедуры оценки в графическом виде:



За описанием процедуры оценки качества следует описание полученных результатов. Для **оценки вклада индивидуальных свойств** и их комбинаций в качество работы системы остальные параметры системы были зафиксированы на значениях по умолчанию: при этом не производится ILP-оптимизация, частотный фильтр на конструкции устанавливается равным 20, соотношение тестовой и тренировочной выборок составляет 40/60. Ниже приводятся пять лучших конфигураций системы, основанных только на синтаксических, только на семантических и на полных наборах свойств.

Табл. 1: ТОП-5 конфигураций системы, все свойства

Features	P	R	F	Acc
Voice,POS,finncase,prep_lemma,case,shortPath4	0.759	0.667	0.695	0.950
Vform,prep_lemma,case,shortPath4	0.765	0.666	0.694	0.951
Vform,finncase,prep_lemma,case,shortPath4	0.764	0.667	0.694	0.950
Voice,vform,POS,finncase,lemma,prep_lemma,path,case,shortPath4	0.754	0.668	0.694	0.950

Vform,finncase,prep_lemma,path,case,shortPath4	0.761	0.667	0.694	0.950
baseline	0.331	0.356	0.343	0.928

Табл. 2: ТОП-5 конфигураций системы, синтаксические свойства

Features	P	R	F	Acc
Vform,prep_lemma,case,shortPath4	0.765	0.666	0.694	0.951
Vform,finncase,prep_lemma,case,shortPath4	0.764	0.667	0.694	0.950
Vform,finncase,prep_lemma,path,case,shortPath4	0.761	0.667	0.694	0.950
Vform,prep_lemma,path,case,shortPath4	0.761	0.667	0.693	0.950
Voice,vform,prep_lemma,case,shortPath4	0.762	0.666	0.692	0.950
baseline	0.331	0.356	0.343	0.928

Табл. 3: ТОП-5 конфигураций системы, семантические свойства

Features	P	R	F	Acc
POS,lemma,cluster-all	0.514	0.424	0.433	0.925
POS,lemma,cluster-nouns	0.514	0.424	0.433	0.925
Lemma,cluster-all	0.513	0.422	0.431	0.925
Lemma,cluster-nouns	0.513	0.422	0.431	0.925
POS,lemma	0.521	0.420	0.429	0.926
baseline	0.331	0.356	0.343	0.928

Формально наилучшие показатели качества достигаются при использовании семантических и синтаксических свойств, однако использование только синтаксических свойств позволяет добиться схожих результатов. В то же время системы, основанные только на семантических свойствах, демонстрируют значительно худшее качество работы, впрочем, всё равно превосходя по качеству базовый классификатор. В главе содержится подробный анализ причин данного поведения системы.

На втором этапе оценки пять лучших конфигураций свойств для каждого из классов свойств были проанализированы с точки зрения **влияния ILP-оптимизации**,

ограничения на частоту встречаемости конструкции и **соотношения тестовой и тренировочной выборок** на итоговое качество работы системы. В ходе тестирования измерялось качество системы на следующих комбинациях параметров:

- ILP: да/нет
- Размер тестовой выборки (tts): 0.1, 0.2, 0.3, 0.4, 0.5
- Ограничение на частоту встречаемости, минимум (thr): 10, 20, 30, 40

Всего было протестировано 600 конфигураций системы. Топ-10 наилучших результатов приводится в следующей таблице:

Табл. 4: ТОП-10 наилучших конфигураций системы (с учетом доп. параметров)

Features + ILP + thr + tts	P	R	F	Acc
vform,finncase,prep_lemma,path,case,shortPath4__ilp_False__thr_40__tts_0.1	0.799	0.744	0.760	0.958
vform,finncase,path,case,shortPath4__ilp_False__thr_40__tts_0.1	0.796	0.744	0.758	0.957
vform,path,case,shortPath4__ilp_False__thr_40__tts_0.1	0.796	0.743	0.757	0.957
vform,finncase,prep_lemma,path,case,shortPath4__ilp_True__thr_40__tts_0.1	0.796	0.733	0.755	0.958
vform,path,case,shortPath4__ilp_True__thr_40__tts_0.1	0.794	0.731	0.752	0.957
vform,finncase,prep_lemma,case,shortPath4__ilp_True__thr_40__tts_0.1	0.799	0.731	0.752	0.959
voice,vform,finncase,path,case,shortPath4__ilp_False__thr_40__tts_0.1	0.793	0.736	0.751	0.956
vform,finncase,prep_lemma,case,shortPath4__ilp_False__thr_40__tts_0.1	0.796	0.735	0.751	0.958
vform,prep_lemma,case,shortPath4__ilp_False__thr_40__tts_0.1	0.795	0.735	0.750	0.957
vform,prep_lemma,case,shortPath4__ilp_True__thr_40__tts_0.1	0.797	0.729	0.750	0.958

Как видно из таблицы, ожидаемо лучшие результаты демонстрируют системы, для которых доступно наибольшее количество тренировочных данных, с набором свойств, показавшим также лучшие результаты на первом этапе тестирования. Несколько неожиданным является тот факт, что ILP-оптимизация не всегда приводит к повышению качества в терминах F-меры. Качество работы систем с большим объёмом тренировочных данных ожидаемо выше независимо от ограничения на частоту встречаемости конструкции. Отметим, однако, что разница в результатах между размерами тестовой выборки 0.2 и 0.5 невелика. Отчасти это объясняется наличием

неточностей в использованных тренировочных и тестовых данных: в случае, когда тестовые данные содержат ошибки разметки, иногда классификатору оказывается "выгодно" иметь большее количество тестовых данных, т.к. таким образом уменьшается процент случайных ошибок, связанных с разметкой и проекцией разметки на синтаксические узлы.

Результаты оценки системы позволяют нам сделать следующие выводы:

1. Наилучшие результаты достигаются при использовании комбинированных семантико-синтаксических наборов свойств, однако и синтаксических свойств зачастую оказывается достаточно для достижения качества, близкого к максимальному. Особое значение имеет свойство "*синтаксический путь от предиката*", которое во многом определяет результат классификации в случаях, когда оно включено в признаковый набор. При этом ограничение длины пути оказывает положительный эффект на качество классификации.

2. Семантические свойства в изоляции показывают менее высокие результаты, однако даже в этом случае качество работы системы превосходит базовый классификатор, выделяющий класс большинства. Интерес вызывает свойство "*кластер*", которое в нашем случае не оказывает почти никакого положительного эффекта на классификацию.

3. Модуль глобальной оптимизации оказывает незначительный положительный эффект на классификацию, и эффект оптимизации наблюдается наиболее отчётливо в случаях, когда исходное качество было невелико.

4. Ограничение на частоту конструкции и увеличение объёма тестовых данных ожидаемо приводят к повышению качества работы системы.

Первые два наблюдения представляют особый интерес и рассматриваются в главе более подробно.

Глава IV
Выводы

Глава IV подводит итоги диссертационного исследования и содержит рекомендации по дальнейшим исследованиям в выбранной области. Проведённый

анализ демонстрирует важность синтаксических свойств для автоматической разметки актантов, а также важность соответствия исходной и целевой предметной областей при использовании дистрибутивных моделей для учёта лексического сходства актантов. Полученные результаты также демонстрируют, что глобальная оптимизация является важным шагом в автоматической обработке актантов.

Наши выводы о возможных альтернативных подходах к решению задачи автоматической разметки актатнов в русском языке можно условно разделить на три группы. Первая группа выводов связана с решениями, которые находятся в русле используемых в работе подходов и так или иначе могли бы способствовать развитию и улучшению разработанной системы. Мы останавливаемся на перспективах использования классификаторов на основе интерпретируемых моделей, на возможных модификациях использованных в работе свойств и усовершенствованиях модуля глобальной оптимизации. Вторая группа выводов касается проблемы использования методов обучения без учителя для решения задач автоматической семантической разметки актантов применительно к русскому материалу. В частности, рассказывается о методе проекции аннотаций, предложенном Х. Фюрстенау[17], который позволяет осуществить проекцию разметки по синтаксическим ролям на неразмеченные данные с помощью выравнивания синтаксических деревьев исходного и целевого предложений. Наконец, финальные замечания связаны с возможными шагами по усовершенствованию корпуса FrameBank. Отмечается важность экспертной синтаксической разметки в подобных ресурсах, проблема репрезентативности данных, а также упоминаются технические возможности по представлению корпуса примеров в одном из стандартных форматов.

Диссертацию завершает **Заключение**, в котором перечисляются основные результаты проделанной работы:

[17] Furstenau, H., & Lapata, M. (2011). Semi-supervised semantic role labeling via structural alignment. Computational Linguistics, 38(1), 135–171

- Разработана и подробно описана система автоматической классификации актантов для русского языка с использованием методов машинного обучения на основе корпуса примеров FrameBank
- Определены группы свойств, использованных для машинного обучения, разработаны принципы интеграции свойств в модуль предобработки данных
- Разработан модуль глобальной оптимизации, контролирующий соответствие результатов работы системы формальным ограничениям теории семантических ролей
- Качество работы системы оценено с использованием изолированных тестовых данных. Определён вклад и особенности отдельных свойств, а также влияние размера выборки и глобальной оптимизации на качество работы системы
- Выявлены потенциальные возможности для оптимизации представленной системы, а также выработаны рекомендации по дальнейшему развитию как представленной системы, так и других подобных систем, основанных на русском материале.

Содержание работы отражено в следующих **публикациях**:

- **Кузнецов И. Автоматическое выделение глагольных актантов: теоретическая основа и актуальные подходы. НТИ. Сер. 2. ИНФОРМ. ПРОЦЕССЫ И СИСТЕМЫ. 2012. No. 12.**
- **Кузнецов И. Автоматическое извлечение двусловных терминов по тематике "Нанотехнологии в медицине" на основе корпусных данных. НТИ. Сер. 2. ИНФОРМ. ПРОЦЕССЫ И СИСТЕМЫ. 2013. No. 5.**
- **Акинина Ю., Кузнецов И., Толдова С. Сравнение двух методов автоматического извлечения участников из неструктурированных источников. НТИ. Сер. 2. ИНФОРМ. ПРОЦЕССЫ И СИСТЕМЫ. 2013. No. 6.**
- Kuznetsov I. Semantic Role Labeling for Russian Language based on Russian FrameBank // Analysis of Images, Social Networks and Texts. 4th International Conference, AIST 2015,

Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers / Ed. by M. Y. Khachay, N. Konstantinova, A. Panchenko, D. I. Ignatov, V. Labunets. Vol. 542: Series: Communications in Computer and Information Science. Springer International Publishing, 2015. P. 337-348.

- И. Кузнецов. Распознавание и классификация актантов в русском языке // Доклады Всероссийской научно-практической конференции «Анализ Изображений, Сетей и Текстов». Екатеринбург, 2013.
- Akinina Y., Kuznetsov I. , Toldova S. The impact of syntactic structure on verb-noun collocation extraction // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). Вып. 12 (19). — М.: Изд-во РГГУ, 2013.