УТВЕРЖДАЮ
Директор ФИЦ ИУ РАН
академик
_____ И.А. Соколов
«25» апреля 2016 г.

О Т З Ы В

ведущей организации «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» на диссертацию Кузнецова Ильи Олеговича на тему «Автоматическая разметка семантических ролей в русском языке», представленную на соискание ученой степени кандидата филологических наук по специальности 10.02.21 – Прикладная и математическая лингвистика.

Диссертационное исследование Кузнецова Ильи Олеговича посвящено очень важной и актуальной теме – автоматической разметке семантических ролей (или актантов) и относится к области создания семантических представлений глагольного распространения для обработки естественно-языковых текстов в информационных системах. Задачи выработки методик и последующей автоматизации семантической разметки текстов на русском языке являются в настоящий момент наиболее приоритетными для области прикладной и математической лингвистики. В центре исследования диссертанта находятся проблемы, которые возникают при создании структурированных лингвистических ресурсов, необходимых для подготовки и разметки текстовых данных для машинного обучения, применяемого в современных интеллектуальных системах обработки знаний, а также для решения широкого класса задач исследования и моделирования естественного языка. Для русского языка подобные ресурсы разработаны еще в очень малой степени.

Актуальность работы

Диссертация Кузнецова Ильи Олеговича «Автоматическая разметка семантических ролей в русском языке» является весьма **актуальной** в настоящий момент. В задачах извлечения знаний из текстов центральной процедурой является установление семантической вершины предложения, клаузы (предиката, выраженного глагольной формой или другой языковой формой, выполняющей роль предиката предложения или клаузы), и корректная идентификация актантов анализируемого предиката является основой правильного извлечения и интерпретации знаний из текста на естественном языке. В работе впервые предложен лингвистически обоснованный способ автоматизации разметки актантов и детально проанализированы результаты её работы, выявлен вклад различных лингвистических свойств и других параметров задачи в качество классификации семантических ролей.

Актуальность диссертационной работы обусловлена стремительно растущей потребностью в обучающих и тестовых корпусах с разметкой по семантическим ролям и доступных инструментах предварительной обработки текста.

В качестве **объекта** исследования диссертант рассматривает проблемы и специфические задачи, возникающие при создании модели классификации актантов на основе деревьев зависимостей и лингвистических свойств изучаемых явлений русского языка.

Работа проводилась на **материале** нового отечественного лингвистического ресурса русского языка FrameBank, содержащего семантически размеченные представления глагольных фреймов, в разработке которого автор диссертации принимает участие, подобное исследование для русского языка ранее не проводилось.

Научная новизна и достоверность исследования

Научная **новизна** диссертационной работы Кузнецова Ильи Олеговича заключается в том, что в ней **впервые** предложены и обоснованы методы

применения систем на основе машинного обучения к корпусу примеров FrameBank, при этом ряд ключевых решений применяется к материалу русского языка впервые.

Результаты исследований верифицированы на основе анализа представительного языкового материала, **достоверно** определены те лингвистические и формальные характеристики, которые необходимы для реализации в системе разметки семантических ролей. Автор диссертационной работы демонстрирует высокую степень владения как лингвистическими, так и вычислительно-алгоритмическими **методами исследования** для решения поставленных задач.

Теоретическая значимость полученных результатов

Диссертация Кузнецова Ильи Олеговича - серьезное самостоятельное исследование. Точно поставлены цели и задачи, результаты исследования убедительно доказаны на представительном фактическом материале. **Теоретическая значимость** диссертации заключается в определении влияния различных лингвистических параметров на качество работы классификатора. Разработанная автором система опирается на такие свойства, как путь в дереве зависимостей, падеж актанта, предлог, которым оформлен актант, лемма и кластер актанта, и другие характеристики. Представленный анализ наглядно демонстрирует важность синтаксических свойств для автоматической разметки актантов в русском языке. Автор убедительно показывает, что роль лексических свойств оказывается второстепенной, и подробно рассматривает возможные причины такого поведения системы.

Практическая значимость полученных результатов

Результаты диссертационного исследования Кузнецова Ильи Олеговича несомненно имеют высокую практическую ценность и технологическое значение и могут быть использованы для создания широкого класса систем обработки естественного, в данном случае русского, языка. Полученные результаты имеют большую **практическую значимость** не только в качестве инструмента лингвистических исследований и создания лингвистических

ресурсов, но и для развития современных информационных технологий. Предложенные автором диссертации методы и реализованные программы являются инновационными и могут быть использованы в системах извлечения знаний из текстов на русском языке и других типов информационных систем, в которых предусмотрены лингвистические процессоры с функциями обучения на текстовых корпусах. Кроме того, полученные результаты имеют большое значение для разработки новых курсов и учебных пособий по математической и прикладной лингвистике.

Для успешного достижения целей диссертационного исследования автором решены поставленные задачи: для обеспечения исходного корпуса FrameBank морфологической и синтаксической информацией автором разработаны и включены в систему специальные доступные ресурсы предобработки; для повышения качества обучающих и тестовых данных произведена фильтрация корпуса примеров FrameBank; разработана модель классификации актантов на основе деревьев зависимостей и лингвистических параметров с учетом характеристик, специфичных для русского языка; разработан модуль глобальной оптимизации, обеспечивающий выполнение ограничений, накладываемых теорией семантических ролей; произведены оценка качества работы полученной системы на отдельной тестовой выборке и оценка влияния лингвистических характеристик и других специальных параметров на качество работы системы; выработаны рекомендации по дальнейшему развитию системы и корпуса FrameBank.

Композиционно диссертация состоит из введения, четырех глав, в которых подробно описаны теоретические и практические результаты, заключения, библиографического списка.

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

Первая глава «Теория семантических ролей и автоматическая разметка актантов» состоит из пяти разделов и посвящена теоретическим основам, истории, начиная с работ, опубликованных в период с 2000-х годов по настоящее время, посвященных системам на основе частично управляемого обучения и неуправляемого обучения. В первой главе также дается характеристика современного состояния концепции семантических ролей для автоматической разметки актантов. Автор диссертационной работы приводит анализ понятия семантической роли, используемого в современной автоматической обработке актантов и основанного на работах Чарльза Филлмора, который ввёл понятие семантического падежа в современную лингвистическую теорию и практику и основал проект FrameNet (лингвистический ресурс, в котором построено системное представление семантических фреймов глаголов английского языка). Автор диссертационной работы опирается на опыт систем автоматической разметки семантических ролей для английского языка с использованием FrameNet, корпуса PropBank и подробно рассматривает работы Дэниэла Журафски, Дэниэла Гилдеа, Марты Палмер и других лингвистов, посвящённые автоматической разметке актантов с использованием семантических ролей; приводит анализ работ Джеффри Грубера, в которых используется понятие тематического отношения, схожее с понятием семантической роли (или семантического падежа).

Во **второй главе** «Система автоматической разметки актантов для русского языка» приводится полное описание системы автоматической разметки актантов для русского языка, разработанной в ходе диссертационного исследования. Глава состоит из пяти разделов, особое внимание уделено описанию параметров и модулей системы разметки. Подробно рассматриваются использованные в системе методы машинного обучения, лингвистические свойства, на основе которых происходит классификация, а также ряд технических решений, использованных при реализации системы и работе с исходными данными.

Третья глава «Экспериментальная оценка и результаты» посвящена экспериментальной оценке качества разработанной системы на стандартных параметрах, по которым можно определить, насколько хорошо работает система. Глава состоит из четырех разделов, в которых подробно описываются процедуры тестирования и глобальной оптимизации системы разметки семантических ролей. В рамках диссертационного исследования оценка качества выполнялась на основании тестовой выборки, также в ряде случаев был произведён экспертный анализ результатов. В третьей главе приведены важные наблюдения о том, что «наилучшие результаты достигаются при использовании комбинированных семантико-синтаксических наборов свойств, однако и синтаксических свойств зачастую оказывается достаточно для достижения качества, близкого к максимальному». Автор указывает, что особое значение имеет свойство “синтаксический путь от предиката”, которое во многом определяет результат классификации в случаях, когда оно включено в набор признаков, при этом ограничение длины пути оказывает положительный эффект на качество классификации.

В **четвертой** главе «Выводы», состоящей из трех разделов, подводятся итоги диссертационного исследования и содержатся рекомендации по дальнейшим исследованиям в рассмотренной области и использованию системы разметки семантических ролей. Автор отмечает, что проведенный 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. Автор указывает, что полученные результаты демонстрируют важность глобальной оптимизации для автоматической обработки актантов.

В **Заключении** суммируются результаты проведенного исследования и намечаются пути развития для направления в целом.

Замечания

Замечания по диссертационной работе сводятся к перечислению замеченных опечаток и стилистических помарок: 1) автор пишет: «…автоматическая разметка актантов – одна из наболее теоретически вовлеченных задач в автоматической обработке языка…» - следует исправить «наболее» на «наиболее», также не очень понятно, что автор подразумевает под «вовлеченностью» применительно к задачам автоматической обработки языка; 2) автор пишет в Заключении « в качестве источника данных был использован корпус примеров FrameNet…», явно имея в виду FrameBank, о котором идет речь в диссертационном исследовании.

Указанные замечания ни в коем случае не снижают высокий научно-теоретический и практический уровень диссертационного исследования и не влияют на общую положительную оценку работы.

Заключение

Таким образом, диссертация Кузнецова Ильи Олеговича представляет собой завершенный научный труд, содержащий существенные результаты научного и практического характера. Материалы исследования опубликованы в 6-ти печатных работах. Диссертационная работа построена методически последовательно, автор демонстрирует высокий уровень лингвистической компетенции, научные положения хорошо аргументированы и достаточно обоснованы. Содержание автореферата полностью отвечает основным положениям диссертации. Тема исследования и полученные результаты соответствуют специальности 10.02.21 – Прикладная и математическая лингвистика.

Диссертационное исследование соответствует всем требованиям, предъявляемым ВАК РФ к диссертациям на соискание ученой степени кандидата филологических наук по специальности 10.02.21 – Прикладная и математическая лингвистика. Диссертация соответствует п. 9, 10 Положения о порядке присуждения ученых степеней и является научно-квалификационной работой, в которой содержится решение задачи, имеющей существенное

значение для соответствующей отрасли знаний, а ее автор Кузнецов Илья Олегович заслуживает присуждения ученой степени кандидата филологических наук по указанной специальности.

Отзыв составлен кандидатом филологических наук Козеренко Еленой Борисовной, заведующей лабораторией Компьютерной лингвистики и когнитивных технологий обработки текстов Федерального исследовательского центра «Информатика и управление» Российской академии наук, 119333, г. Москва, ул. Вавилова 44 корпус 2, www.ipiran.ru. Статьи Козеренко Е.Б.:

Козеренко Е.Б. Интегральное моделирование языковых структур в лингвистических процессорах систем обработки знаний и машинного перевода // Информатика и ее применения, 2014. Т. 8. Вып. 1. С. 89-98.

Козеренко Е.Б. Стратегии выравнивания параллельных текстов: семантические аспекты // Информатика и ее применения, 2013. Т. 7. Вып. 1. С.82-89.

Kozerenko, E. Functional and Cognitive Aspects in Linguistic Modelling // Proceedings of ICAI'13, WORLDCOMP'13, July 22-25, 2013, Las Vegas, Nevada, USA. CRSEA Press, USA, Vol. II, 2013. P. 896-902.

Отзыв обсужден и утвержден на заседании Лаборатории «Компьютерной лингвистики и когнитивных технологий обработки текстов» ФИЦ ИУ РАН – ИПИ РАН 19 апреля 2016 г., протокол № 1.

Заведующий лабораторией
«Компьютерной лингвистики и когнитивных
технологий обработки текстов»
кандидат филологических наук Е.Б. Козеренко

« 19 » апреля 2016 г.

Подпись Е.Б. Козеренко заверяю
Ученый секретарь ФИЦ ИУ РАН
доктор технических наук Захаров В.Н.

~~Подпись заверяю~~

20.04.2016