ОТЗЫВ

официального оппонента
о диссертации И. О. Кузнецова
“Автоматическая разметка семантических ролей в русском языке”,
представленной на соискание ученой степени кандидата филологических наук
по специальности 10.02.21 – Прикладная и математическая лингвистика.

Диссертация И. О. Кузнецова посвящена одной из задач автоматической обработки текстов на русском языке. Речь идет о выделении семантических ролей – актуальной и далекой от решения задачи. Актуальность задачи не вызывает сомнений, т.к. разметка семантических ролей позволит перейти к более глубокому уровню понимания текста компьютером. В виду большой сложности семантического анализа это направление мало исследовано даже для английского языка и практически совсем не исследовано для русского.

Диссертант предложил подход и разработал систему автоматической разметки актантов и маркировки их семантических ролей для русского языка. Подход основан на машинном обучении с учителем и опирается на ряд лингвистически мотивированных признаков для описания объектов. В качестве метода машинного обучения использован метод опорных векторов, хорошо зарекомендовавший себя для широкого круга задач. В качестве обучающего и тестового корпуса выбран FrameBank. Диссертант квалифицированно применяет различные инструменты автоматического лингвистического анализа, в частности, предназначенные для предобработки текста, синтаксического анализа. Все принимаемые решения тщательно обосновываются, альтернативные варианты обсуждаются. Правильный выбор технических решений, обусловленных различными сложными факторами, свидетельствует о высоком уровне квалификации И. О. Кузнецова в области компьютерной лингвистики.

Для улучшения качества работы системы диссертантом разработан модуль глобальной оптимизации. Созданная система тщательно протестирована, проведена большая экспериментальная работа по оценке вклада различных обучающих параметров в итоговый результат. Интересность, полезность и высокий уровень данной диссертации не вызывает сомнений.

Перейдем к замечаниям и обсуждению спорных моментов. Прежде всего, обращает на себя внимание большое число описок, так на стр. 108 два ошибки встречаются даже в пределах одной фразы: “Зона, отмеченная на Рисунок 344 …”. Далее, кажется, что в диссертации неоправданно много места выделяется для популярного изложения хорошо известных понятий (например, задачи

классификации, рис. 2 и далее). Содержательно, принципиальными являются следующие замечания.

1. При применении методов машинного обучения традиционно используются лингвистические признаки, обладающие следующими свойствами: однозначная интерпретация экспертами, высокая точность распознавания (близкая к 100%), наличие признанных общедоступных стандартных программ выделения признаков. Например, свойство “начинаться с заглавной буквы”. Свойство “путь”, используемое в работе, этими свойствами не обладает. Нет однозначной интерпретации синтаксической структуры предложений, набора синтаксических отношений. Используемая для определения путей программа Шарова вряд ли может считаться стандартной и общедоступной. Точность определения вершин дерева у нее лишь 82%. Точность выделения путей, вероятно, падает с увеличением длины пути. Таким образом, использование этого признака противоречит обычной практике машинного обучения, снижает воспроизводимость результатов диссертации и делает их слишком зависимыми от конкретного выбранного формализма. В то же время семантические роли шире конкретных синтаксических формализмов.

2. Результаты, полученные в диссертации, демонстрируют определяющую роль именно этого признака. В итоге, программа обучается, фактически, нахождению определенных актантных позиций в синтаксическом дереве, а не семантических ролей, как это должно быть согласно названию диссертации. В примерах 14 и 15 (стр. 147) семантическая роль слова “Иван”, очевидно, одна и та же, однако, в дереве синтаксического разбора они занимают разные позиции, что оказывает влияние на результаты распознавания.

3. Странно, что среди лингвистических признаков нет такого важного и часто используемого, как порядок слов. Причем это никак не обсуждается. Без него непонятно как можно правильно определить семантические роли актантов в предложениях типа: *“Спартак” переиграл “Зенит”*. Являясь, формально, двусмысленным, реально оно интерпретируется однозначно.

4. Недостаточно внимания уделено признаку, на который автор, по его же словам, возлагал надежды – кластерам. Стоило поэкспериментировать с кластерами, полученными на основе другого (большего по объему или близкому по тематике) набора данных.

Несмотря на сделанные замечания, частично полемического характера, я высоко оцениваю данную диссертацию.

Автореферат и публикации соответствуют содержанию диссертации. Диссертация является научно-квалификационной работой, решившей задачу,

актуальную для прикладной лингвистики. На основании параграфов 9 и 10 Положения о присуждении ученых степеней можно утверждать, что И. О. Кузнецов заслуживает присуждения ему ученой степени кандидата филологических наук по заявленной специальности.

Доктор физико-математических наук, профессор,
ведущий научный сотрудник
НОЦ по лингвистике им. И. А. Бодуэна де Куртенэ
ФГАУВО "Казанский (Приволжский) федеральный университет"
Соловьев В.Д.

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАУ ВО «КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
ОГРН 1021602841391 УПРАВЛЕНИЕ ДОКУМЕНТООБОРОТА И КОНТРОЛЯ
ПОДПИСЬ 11.04.2016
Соловьева В.Д. заверяю
Документовед Эдуонова И.Н.

Сведения об оппоненте:

Соловьев Валерий Дмитриевич, maki.solovyev@mail.ru, +79196910489
Доктор физико-математических наук, профессор,
ведущий научный сотрудник
НОЦ по лингвистике им. И. А. Бодуэна де Куртенэ
ФГАУВО "Казанский (Приволжский) федеральный университет" (Казань, 420008, ул. Кремлевская, д.18), public.mail@kpfu.ru

Публикации оппонента:

Valery Solovyev and Vladimir Ivanov. Knowledge-driven Event Extraction in Russian: Corpus-based Linguistic Resources. Computational intelligence and neuroscience. 2016, Volume 2016, Article ID 4183760, http://www.hindawi.com/journals/cin/aip/698102/

Solovyev V., Kibrik A. How can computer technologies help linguistic typology? Herald of the Russian Academy of Sciences. 2015. V. 85, Issue 1, pp 33-39.

Elizarov A. M., Lipachev E. K., Nevzorova O. A., and Solov'ev V. D. Methods and Means for Semantic Structuring of Electronic Mathematical Documents. Doklady Mathematics, 2014, Vol. 90, No. 1, pp. 521–524.