

«УТВЕРЖДАЮ»

ВРИО директора ИППИ РАН

Доктор физико-математических

наук, профессор РАН

А.Н. Соболевский

« » 2016г.



ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

Федерального государственного бюджетного учреждения науки

«Институт проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН)»

на диссертацию КУДИНОВА М.С.

«Статистическое моделирование русского языка с помощью нейронных сетей»,
представленную на соискание ученой степени кандидата технических наук по
специальности 05.13.17 – теоретические основы информатики.

Актуальность. Тема диссертации связана с определением характеристик естественного языка. Рассматривается задача построения статистической модели целевого языка, т.е. модели условного распределения следующей в тексте словоформы по нескольким известным предыдущим словам. Статистические модели языка используются в распознавании речи, распознавании рукописного текста и системах предиктивного набора текста на мобильных устройствах. Наиболее популярный подход к моделированию языка состоит в использовании n-граммных моделей. Этот подход обладает определенными недостатками, которые пытаются компенсировать методами искусственных нейронных сетей. Актуальность исследования подтверждается также внедрением результатов работы в сферу профессиональной деятельности компаний, работающих в отрасли информационных технологий, в частности Samsung Electronics Ltd.

Научная новизна и теоретическая значимость. В рамках нейросетевых подходов задача языкового моделирования сводится к задаче классификации, где каждый класс соответствует слову в заданном словаре. Сложность алгоритма предсказания по времени и

по памяти при этом линейны по размеру словаря. Использование нейронных сетей позволило существенно улучшить показатели перплексии на стандартных англоязычных тестовых корпусах, однако применение таких моделей к языкам с развитой морфологией существенно осложняется большей разреженностью данных и меньшей вычислительной эффективностью. Другой открытой проблемой языкового моделирования является проблема учета в предсказании слов, находящихся на дальнем расстоянии от прогнозируемого. В работе М.С.Кудинова предлагаются подходы к решению двух перечисленных проблем. Проблему разреженности предлагается решать за счет перехода от прогнозирования последующей словоформы к раздельному предсказанию начальных форм слов и их грамматических показателей (например, числа и падежа). Вторая проблема решается автором путем расширения рекуррентной архитектуры за счет признаков, полученных из модели вероятностного тематического моделирования. Предполагается, что модель тематического моделирования способна дать эффективное сжатое представление всего доступного левого контекста и тем самым учитывать зависимости между словами на произвольном расстоянии. Предложенные автором подходы проверяются экспериментально. Решение поставленных автором задач имеет большое значение для методологии распознавания речи и анализа текста.

Основные результаты. Первая глава «Задача статистического моделирования языка» носит обзорный характер. В ней дается формальная постановка задачи статистического моделирования языка и приводится анализ существующих подходов к ее решению и используемых метрик качества.

Во **второй главе** «Моделирование языка с помощью рекуррентных нейронных сетей» приводятся экспериментальные данные, указывающие на то, что использование языковой модели русского языка, основанной на рекуррентной нейронной сети, оказывается эффективнее в смысле перплексии и уменьшения пословной ошибки в системе распознавания речи по сравнению со сглаженной 5-граммной моделью Кнесера-Нея только в том случае, если модель предсказывает только начальные формы слов. Языковые модели были обучены на новостном корпусе объемом 2 млн. словоупотреблений. Для оценки уменьшения пословной ошибки проводился эксперимент по переранжированию готового списка гипотез. Основываясь на результатах эксперимента, автор предлагает подход, основанный на раздельном предсказании начальных форм слов (лемм) и их грамматических показателей. Это позволяет на порядок уменьшить количество классов при предсказании, тем самым увеличивая производительность модели. С другой стороны

раздельное предсказание начальных форм и грамматических форм снижает разреженность данных.

В **третьей главе** «Расширение рекуррентной архитектуры с помощью тематического моделирования» автором предлагается оригинальная языковая модель, целью которой является учет дальних зависимостей между словами в тексте. Подход автора к проблеме дальних зависимостей основан на расширении рекуррентной нейронной сети за счет признаков, полученных из тематического разложения левого контекста предсказываемого слова в рамках теории вероятностного тематического моделирования, развиваемой К.В.Воронцовым. Автором проанализирован эффект, который оказывает выбор метода аддитивной регуляризации тематической модели на эффективность статистической языковой модели в смысле перплексии. Эксперименты, поставленные автором демонстрируют, что предлагаемая автором модель оказывается более эффективной в смысле перплексии, чем исходная рекуррентная архитектура.

В **четвертой главе** «Предсказание морфологических характеристик с помощью нейронной сети» автором предложен оригинальный подход к предсказанию грамматических признаков леммы с помощью сверточной нейронной сети. Утверждается, что в силу специфики грамматики языка при прогнозировании грамматической формы проблема учета дальних зависимостей отсутствует, следовательно, нет необходимости в рекуррентной архитектуре. Выбор сверточной архитектуры слоев продиктован главным образом соображениями производительности. Эффективность модели проверяется в эксперименте по переранжированию гипотез, возвращаемых системой распознавания речи. Результаты указывают на то, что использование модели, предлагаемой автором, приводит к уменьшению пословной ошибки.

Практическая значимость результатов, полученных в диссертации, заключается в устойчивости разработанного метода к порядку слов в произвольной речи на русском языке. Эти результаты могут быть использованы при разработке программных продуктов в области обработки естественного языка. М.С.Кудинов является автором патента RU 2583150 «Голосовая связь на естественном языке между человеком и устройством», работа над которым велась в рамках диссертационного исследования. Результаты, полученные М.С.Кудиновым в рамках диссертационного исследования, используются в разработках компании Samsung Electronics Ltd. Результаты диссертации, алгоритмическое и программное обеспечение, разработанное автором считаем возможным рекомендовать к рассмотрению и использованию в заинтересованных организациях РАН (ФИЦ

"Информатика и управление", ФГБЮН СПИИ РАН, ИППИ РАН), других ведомствах (в/ч 35533), ООО "ЦРТ", ООО "СТЭЛ-КС".

Замечания. Отмечая актуальность, новизну и значимость диссертационного исследования, следует высказать следующие замечания:

1. Несмотря на то, что в главе 2 автор приводит формальное доказательство невозможности моделирования дальних зависимостей с помощью стандартной рекуррентной нейронной сети, автор далее не приводит теоретических оценок того, насколько эффективен подход, описанный в главе 3. Данное утверждение справедливо и для главы 4. Поэтому в качестве основного недостатка работы отметим недостаточность теоретических обоснований.
2. Результаты, приведенные в главе 4, в целом не позволяют сделать однозначный вывод об эффективности рекуррентной модели, поскольку получено лишь незначительное снижение вероятности ошибки распознавания слов. Приводимые автором доводы указывают на то, что эксперименты необходимо продолжить, и результаты главы 4 носят скорее предварительный характер.
3. К недостаткам работы следует отнести использование автором обучающих и тестовых данных, не находящихся в открытом доступе, что затрудняет воспроизведение результатов, полученных автором.
4. В главы 2 и 3 включено описание предшествующих методов, которые должны были быть описаны в обзорной главе 1.

Указанные замечания не отменяют положительную оценку выполненной автором работы. Диссертация логично построена, ее структура и содержание соответствуют цели и задачам исследования. Автореферат достаточно полно отражает содержание диссертации. Основные результаты опубликованы в 7 печатных работах, из них 5 в рецензируемых изданиях.

Диссертация Кудинова Михаила Сергеевича «Статистическое моделирование русского языка с помощью нейронных сетей» является самостоятельной законченной научно-квалификационной работой. Диссертация обладает научно-теоретической и практической значимостью, ее результаты имеют существенное значение для развития направлений исследований в области теоретических основ информатики, а именно, в исследовании моделей и алгоритмов анализа данных, алгоритмов анализа текста, устной речи и изображений.

Работа Кудинова Михаила Сергеевича «Статистическое моделирование русского языка с помощью нейронных сетей» отвечает всем требованиям пп. 9–14 Положения о присуждении ученых степеней, предъявляемым к кандидатским диссертациям, а ее автор

заслуживает присвоения ученой степени кандидата технических наук по специальности 05.13.17 – теоретические основы информатики.

Отзыв выполнен ведущим научным сотрудником лаборатории Информационных технологий передачи, анализа и защиты информации, доктором физико-математических наук, доцентом Сорокиным Виктором Николаевичем, обсужден и одобрен на заседании лаборатории №3 Информационные технологии передачи, анализа и защиты информации, протокол №3 от 21.10.2016.

Ведущий научный сотрудник лаборатории
Информационных технологий передачи,
анализа и защиты информации, доктор
физико-математических наук, доцент

В.Н.Сорокин



e-mail: vns@iitp.ru,

тел.: 8-495-699-50-96