

ОТЗЫВ

официального оппонента к.т.н. Ромашкина Юрия Николаевича
на диссертацию Кудинова Михаила Сергеевича
на тему "Статистическое моделирование русского языка с помощью нейронных
сетей", представленную на соискание ученой степени
кандидата технических наук по специальности
05.13.17 – "Теоретические основы информатики"

Речевые технологии активно развиваются в настоящее время и находят все более широкое применение в различных сферах деятельности человека, в том числе в задачах автоматического распознавания слитной речи и рукописного текста, тематической классификации текстов. Одной из важнейших задач при разработке этих технологий является создание адекватной модели конкретного языка. Существующие подходы к ее решению используют основные положения теории вероятности применительно к аппроксимации существующих зависимостей между словами в их последовательности. К настоящему времени разработано и апробировано множество различных алгоритмов статического моделирования языка, обеспечивающих удовлетворительное качество работы в некоторых практических приложениях преимущественно для ряда иностранных языков. Вместе с тем развитие и внедрение этих алгоритмов для приложений на русском языке заметно сдерживается как по причине отсутствия свободного доступа к необходимым речевым корпусам, так и вследствие особенностей русского языка, связанных с произвольным порядком слов в предложениях. Поэтому тему диссертационной работы Кудинова М.С. можно считать актуальной.

Целью рассматриваемой диссертационной работы является дальнейшее совершенствование существующих подходов к статистическому моделированию русского языка на базе современных достижений в развитии нейросетевых алгоритмов обработки информации.

В диссертации решены следующие основные научно-технические задачи:

- выполнены анализ ранее разработанных методов и систематизация полученных знаний при статистическом моделировании различных языков, по результатам которых определены их основные преимущества и недостатки, сформулированы перспективные пути повышения качества их работы применительно к русскому языку, обладающему высокой флективностью;

– предложены и программно реализованы новые статистические модели русского языка, использующие рекуррентные и сверточные нейронные сети, методы тематического моделирования, классификации и предсказания морфологических форм слов;

– получены экспериментальные оценки показателей качества работы предложенных моделей русского языка, доказывающие их определенные преимущества и показывающие направления дальнейшего совершенствования.

В ходе решения указанных задач соискателем получены результаты, обладающие научной новизной и практической значимостью.

Первая глава диссертационной работы посвящена обзору и анализу существующих методов статистического моделирования языка. В ней с пониманием излагаются вероятностные основы такого моделирования, характеризующие распределения различных конечных последовательностей слов в языке. Приводятся наиболее широко применяемые показатели качества языковой модели, справедливо указывается, что каждая из них в отдельности не является достаточной. По материалам научно-технических публикаций проводятся обзор и анализ наиболее известных методов статистического моделирования языка, достоверно определены их основные преимущества и недостатки, задающие направления диссертационных исследований соискателя.

Во второй главе излагаются общие принципы моделирования языка с помощью нейронных сетей и, в частности, рекуррентной сети Элмана. Приводятся математические основы метода их обучения с применением алгоритма обратного распространения ошибки. Верно указан главный недостаток этого метода, ограничивающий возможность учета влияния дальних предшествующих слов в последовательности. Предлагается архитектура рекуррентной нейросети с внешним классификатором, снижающая сложность процедуры ее обучения. Экспериментально подтверждается эффективность модели русского языка на базе рекуррентной нейросети по показателям перплексии и ранжирования гипотез распознавания.

Способ учета длинного контента в гибридной рекуррентной нейросети с дополнительным вероятностным тематическим моделированием исследуется соискателем в третьей главе. Устанавливается, что задача выбора оптимальных параметров тематической модели сводится к задаче матричного разложения, минимизирующего взвешенную сумму дивергенций Кульбака-Лейблера. Для повышения устойчивости алгоритма обучения рассматриваются методы с

различными аддитивными регуляризаторами. Описываются проведенные эксперименты, подтверждающие, что предложенная гибридная архитектура обеспечивает примерно 10 % снижение перплексии относительно исходной рекуррентной модели.

В четвертой главе диссертационной работы обосновывается целесообразность применения сверточной нейросети для предсказания морфологических характеристик леммы. Преимущество такой нейросети связано также с заметным повышением скорости ее обучения, позволяя более точно подбирать значения параметров и в итоге повышать качество языковой модели. Излагаются результаты экспериментов, показывающие повышение качества языковой модели по показателю кросс-энтропии для каждой изменяемой части речи русского языка. Экспериментально подтверждается увеличение точности автоматического распознавания русской речи с использованием комбинации предложенной сверточной нейросети и 3-граммной модели со сглаживанием Кнесера-Нея.

Обоснованность сформулированных научных положений и достоверность полученных результатов подтверждается корректным использованием математического аппарата и экспериментальной проверкой предложенных способов повышения качества языковых моделей.

Вместе с тем диссертационная работа имеет следующие недостатки.

1. При изложении материала исследований соискателю не удалось избежать орфографических и синтаксических ошибок, а также расхождений в приводимых численных данных. Имеются незначительные отклонения от требований ГОСТ 2.105 к оформлению текстовых документов.

2. В разделе 1.4.5 используются параметры S и S_j , физический смысл которых напрямую не указан.

3. В разделе 2.7.2 появляется новый показатель – точность выбора правильной гипотезы (SER), который отсутствует в разделе 1.3 при описании методов оценки качества языковой модели.

4. Соискатель неоправданно использует в качестве синонимов понятия "мера" и "метрика", имеющие разные математические определения.

В целом, несмотря на указанные выше замечания, диссертационная работа Кудинова М.С. представляет собой законченное научное исследование, содержащее ряд новых решений применительно к статистическому моделированию русского языка.

Материал диссертации подается соискателем в логической последовательности, продиктованной поставленной целью работы, и раскрывает решаемые научно-технические задачи. В ней содержится достаточное для понимания количество иллюстраций и экспериментальных данных. Аппарат теории вероятностей используется корректно.

Основные результаты работы представлены в рецензируемых научно-технических журналах и в виде тезиса доклада на международной конференции. Содержание автореферата соответствует содержанию диссертации.

Диссертация выполнена в соответствии требованиями п.9 Положения о порядке присуждения ученых степеней и является научно-квалифицированной работой, в которой содержится решение задачи, позволяющее повысить качество разработки и эффективность применения современных речевых технологий.

Считаю, что Кудинов М.С. заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.17 – "Теоретические основы информатики".

Официальный оппонент к.т.н.

Федеральное государственное
казенное учреждение

"Войсковая часть 35533"

143988. г. Железнодорожный Московской области

Телефон 8-495-527-01-17.

E-mail romayn@yandex.ru



Ю.Н. Ромашкин

Подпись Ромашкина Ю.Н. удостоверяю

Ученый секретарь диссертационного совета

5 декабря 2016 г.



Н.В. Боровик