На правах рукописи

УДК 004.852

Кудинов Михаил Сергеевич

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ РУССКОГО ЯЗЫКА С ПОМОЩЬЮ НЕЙРОННЫХ СЕТЕЙ

Специальность 05.13.17 —
«Теоретические основы информатики»

Автореферат диссертации на соискание учёной степени
кандидата технических наук

Москва — 2016

Работа выполнена в Федеральном исследовательском центре «Информатика и управление» Российской Академии Наук (ФИЦ ИУ РАН)

Научный руководитель:

Кандидат физико-математических наук, ведущий научный сотрудник — **Чучупал Владимир Яковлевич**

Официальные оппоненты:

Кандидат технических наук, ведущий научный сотрудник
Федеральное государственное казенное учреждение "Войсковая часть 35533" — **Ромашкин Юрий Николаевич**

Доктор технических наук, доцент
заведующий лабораторией речевых и многомодальных интерфейсов
Федеральное государственное бюджетное учреждение науки Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН) — **Карпов Алексей Анатольевич**

Ведущая организация: Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН)

Защита состоится _____ 2016 г. в _____ час. _____ мин. на заседании диссертационного совета Д002.073.05 при Федеральном государственном учреждении «Федеральный исследовательский центр Информатика и управление» Российской академии наук, по адресу: 119333, г. Москва, ул. Вавилова, д. 40.

С диссертацией можно ознакомиться в библиотеке и на сайте ФИЦ ИУ РАН – http://www.frccsc.ru/diss-council/00207305/diss

Автореферат разослан _____ 2016 г.

Ученый секретарь
диссертационного совета,
д.ф.-м.н., профессор _____ Рязанов В.В.

Общая характеристика работы

Актуальность темы Задача статистического моделирования языка состоит в определении вероятностного распределения над цепочками слов в некотором языке. Данная задача естественным образом возникает в таких практических областях как распознавание речи, оптическое распознавание символов (OCR), распознавание рукописного текста, машинный перевод, проверка орфографии, предикативный ввод и других.

За минувшие 25 лет спрос на программные решения, связанные с обработкой текста, уже неоднократно переживал периоды роста, связанные сначала с появлением персональных компьютеров, затем со стремительным развитием интернета, и, наконец, с одновременным взрывным ростом социальных сетей и рынка мобильных устройств. При этом естественный язык остается важнейшим способом коммуникации, будь то ввод поискового запроса на миниатюрном экране мобильного телефона, подсказки автомобильного навигатора или бизнес-переписка. Практически во всех таких приложениях так или иначе используется языковая модель. Так, для удобного ввода текстов на мобильном телефоне, необходимо использовать системы предиктивного ввода, что практически сводится к прямому применению языковой модели; языковая модель — неотъемлемая часть систем распознавания речи, в том числе и в голосовом поиске; языковые модели используются в системах машинного перевода, качество которых на настоящий момент еще далеко от идеального, но все же неуклонно растет.

Историю языкового моделирования принято возводить к работам Шеннона, однако настоящая популярность статистических методов обработки текста началась лишь в 1980-е, с первыми успехами, полученными инженерами компании IBM. За сравнительно недолгую историю существования проблемы статистического моделирования языка было предложено большое количество различных подходов к ее решению, главным из которых по сей день остается подход, основанный на сглаженных n-граммных моделях. Было экспериментально продемонстрировано, что преимущества более сложных моделей в целом исчезают с ростом обучающей выборки. С начала 2010-х стал стремительно развиваться подход, основанный на рекуррентных нейронных сетях, описанный Т.Миколовым. В его работах было продемонстрировано, что с ростом объема обучающего корпуса преимущества предложенной им нейросетевой модели только увеличиваются. Рекуррентные модели за прошедшие годы нашли применение в различных областях от диалоговых систем до генерации текста по изображению.

Попытка прямолинейного применения модели Миколова к флективным языкам сталкивается с проблемой разреженности данных и большой вычислительной сложности в связи со свободным порядком слов и большим количеством грамматических форм, свойственным флективным языкам, в частности русскому. Одним из решений данной проблемы является раздельное предсказание начальных форм слова (лемм) и их морфологических форм. Данный подход позволил бы существенно снизить вычислительные затраты для обучения статистической модели русского языка.

Целью данной работы является построение эффективной статистической модели русского языка.

Для достижения поставленной цели необходимо было решить следующие задачи:

1. Аналитический обзор состояния проблемы и систематизация подходов к статистическому моделированию языка;
2. Экспериментальная оценка качества существующих языковых моделей для русского языка, выявление их недостатков и способов их устранения;
3. Разработка и теоретическое описание новых модификаций языковых моделей, устраняющих выявленные недостатки;
4. Разработка алгоритмов и программная реализация полученных моделей, а также экспериментальная проверка их эффективности.

Основные положения, выносимые на защиту:

1. Использование рекуррентной нейронной сети для статистического моделирования русского языка для предсказания начальных форм слова (лемм) более эффективно, чем n-граммная языковая модель, как с вычислительной точки зрения, так и с точки зрения качества предсказания слов (уменьшения перплексии тестовых данных);
2. Предсказание словоформ с помощью рекуррентной нейронной сети является неэффективным с вычислительной точки зрения;
3. Языковая модель, использующая отдельные классификаторы для предсказания лемм и морфологических признаков, требует значительно меньших вычислительных затрат при той же перплексии;
4. Расширение рекуррентной модели на леммах за счет добавления признаков, полученных путем отображения левого контекста текущего слова в вектор действительных чисел, приводит к улучшению показателя перплексии новой модели по сравнению с исходной;
5. Предсказание морфологической формы, реализованное с помощью сверточной нейронной сети, задействующей морфологическую и лексическую информацию, снижает процент пословной ошибки при распознавании речи.

Научная новизна:

1. Впервые предложена статистическая языковая модель с раздельным предсказанием лемм и морфологических признаков, основанная на применении рекуррентной и сверточной нейронных сетей, позволяющая учитывать более длинный левый контекст для предсказания;

2. Разработана и реализована гибридная статистическая языковая модель на рекуррентной нейронной сети и тематическом разложении левого контекста, повышающая качество на более длинных текстах;

3. Впервые предложен, обоснован и экспериментально исследован метод расширения словаря языковой модели на рекуррентной нейронной сети за счет отдельной модели предсказания морфологической формы русских слов, основанной на сверточной нейронной сети.

В диссертации показана возможность построения статистической языковой модели, основанной на нейросетевом подходе, обеспечивающей эффективное предсказание морфологических форм и способной к учету дальних контекстных зависимостей между словами. Это определяет **теоретическую ценность** работы.

Использование предложенной модели в системах распознавания речи и текстового ввода позволяет улучшить качество данных систем. Этим определяется **практическая ценность** работы. Работы автора нашли практическое применение в технологиях компании «Самсунг Электроникс Ко., Лтд». В частности, автором был получен патент на изобретение «Голосовая связь на естественном языке между человеком и устройством» (RU 2583150).

Степень достоверности полученных результатов обеспечивается сходимостью теоретических оценок и экспериментальными результатами.

Работа проходила **апробацию** Основные результаты работы докладывались на: конференциях «Диалог–2014» (Бекасово, 2014), «Диалог–2015» (Москва, 2015), «SPECOM-2015» (Афины, 2015); семинаре Вычислительного центра им. Дородницына ФИЦ ИУ РАН.

Все теоретические и экспериментальные результаты получены автором **лично**.

Публикации. По тематике исследований опубликовано 7 научных работ, в том числе 5 статей в журналах, рекомендованных ВАК.

Содержание работы

Во **введении** обоснована актуальность работы, ее научная и практическая ценность, кратко сформулированы основные цели и задачи работы, представлены основные положения, выносимые на защиту.

Первая глава посвящена обзору методов статистического моделирования языка.

Основными метриками качества в оценке языковых моделей являются перплексия и уровень пословной ошибки (для распознавания речи). Ни один из этих показателей не является исчерпывающим: недостатком перплексии является предположение о полной и истинной информации о левом контексте, что не соответствует действительности для задачи распознавания речи. Снижение уровня пословной ошибки, в свою очередь, зависит от изначальной конфигурации системы распознавания, что делает результаты, полученные разными исследовательскими группами, несравнимыми.

Сглаженная n-граммная модель работает достаточно хорошо, причем с ростом объема обучающей выборки перплексия, как и уровень пословной ошибки, падают.

Применение модели, допускающей **перестановки**, не является оправданным даже для флективных языков, так как возможность перестановки слов в языках со свободным порядком слов все же довольно ограничена.

Среди моделей, не использующих векторное представление словаря, наиболее эффективными являются **модели с кэшированием**, однако применение кэширования для распознавания речи затруднено ввиду проблемы «закрепления ошибки», когда неверное распознавания одного слова, существенно влияет на качество распознавания последующих слов. То же верно для модели **максимальной энтропии** в случае, когда используются признаки, основанные на триггерных словах.

Более сложные модели оказываются полезны прежде всего для борьбы с разреженностью данных. С ростом объема обучающей выборки для большинства техник наблюдается постепенное уменьшение разности в перплексии сглаженных n-граммных моделей и их интерполяции с моделями, использующими дополнительную информацию.

Факторная модель по сути представляет собой унифицированный способ использования различной статистической информации в рамках общей генеративной модели. В этом смысле факторная модель наследует как преимущества, так и недостатки всех моделей, которые она включает в себя.

Эффективными оказываются модели, основанные на **распределенном представлении словаря**, т.е. те модели, в которых каждая словоформа отображается в n-мерный вектор. К таким моделям относятся языковые модели, основанные на латентном семантическом анализе, латентном размещении Дирихле, нейронных сетях и тематических моделях.

Использование специальных **лингвистических данных** оправданно для языков с богатой морфологией и при сравнительно небольших объемах обучающих данных, когда невозможно обеспечить корректное обучение для словаря объемом почти в 10 раз большего, чем в случае английского языка.

Наилучшие результаты были достигнуты при помощи моделей, использующих **рекурентные нейронные сети**. Важнейшим преимуществом таких моделей является распределенное представление словаря, однако обучение нейронных сетей представляет собой отдельную достаточно сложную задачу.

Комбинации наиболее эффективных моделей дают результаты, как минимум не хуже, чем каждая из техник в отдельности. Эффективный способ комбинирования является важной отдельной задачей.

В отсутствие больших обучающих корпусов представляется логичным **комбинирование** некоторой сложной модели, основанной на применении лингвистической информации и векторных представлениях, полученных применением тематических моделей и нейронных сетей.

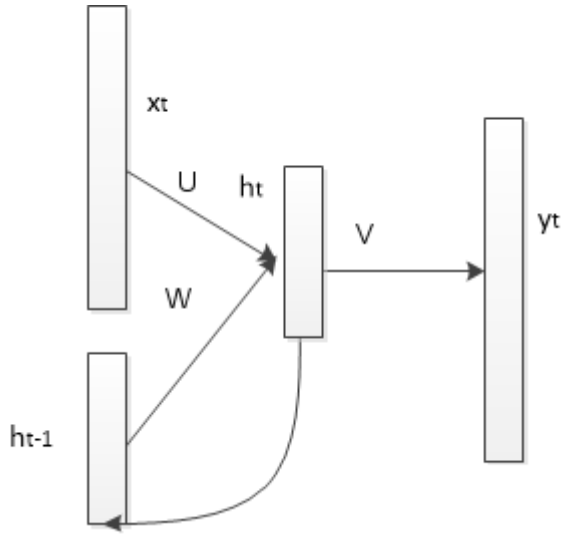

Рисунок 1: Общий вид рекурренной сети Элмана. x_t — входной слой на шаге t; h_t — скрытый слой; y_t — выходной слой.

Во второй главе рассматриваются вопросы, связанные с применением рекуррентных нейронных сетей для построения языковых моделей.

В частности, рассмотрен метод векторного представления слов при помощи нейронных сетей. Отображение осуществляется путем умножения матрицы вложения U на вектор $x = \delta_{w_i}$, содержащий единственную (ненулевую) единичную координату с индексом i, соответствующим индексу слова в словаре:

$$C(w_i) = U \cdot x.$$

Данный метод лежит в основе большинства нейросетевых алгоритмов для работы с естественным языком. В частности, рекуррентной нейронной сети (рис.1). Ее основной отличительной чертой является то, что в вычислении выхода сети на шаге t используется предыдущее значение скрытого слоя с задержкой на один такт. Таким образом, каждый элемент на шаге t связан с каждым элементом скрытого слоя на шаге $t-1$. Пусть даны два зависимых временных ряда $x(t)$, $y(t)$. Тогда рекуррентная нейронная сеть есть функция, аппроксимирующая условное распределение $\mathsf{P}(y^{(t)}|x^{(t)})$ согласно формулам:

$$h_t = f(W \cdot h_{t-1} + U \cdot x_t + b)$$

$$y_t = g(V \cdot h_t + d),$$

где W,U,V — матрицы весов, b, d — смещения, $x \in \mathbb{U}$ — элемент предикторной последовательности на шаге t, $y \in \mathbb{T}$ — распределение вероятностей элементов неизвестной последовательности на том же шаге t, $h \in \mathbb{H}$ — скрытый слой сети, f и g — функции активации. В задаче языкового моделирования в качестве распределения y_t выбирается $\mathsf{P}(w_t|w_t \dots w_1)$. Входной элемент x_t представляется единичным вектором δ_{w_i}, как описано выше.

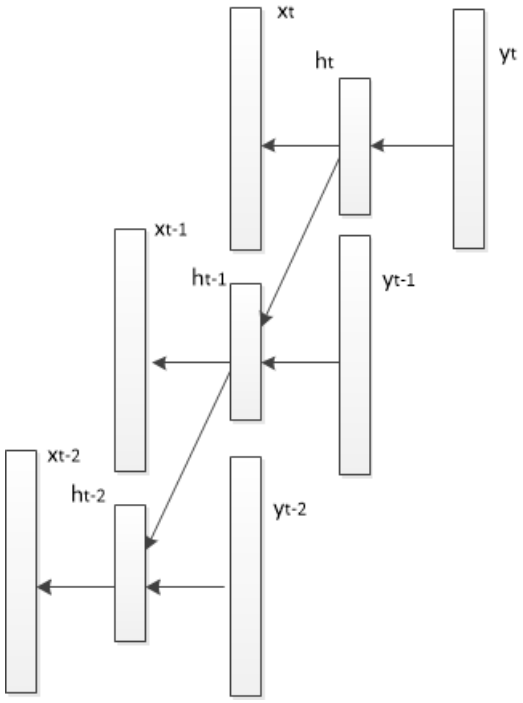

Рисунок 2: Схема вычислений в алгоритме распространения ошибки обратно по времени. y_k — выходной слой на шаге t; x_t входной слой; h_t — скрытый слой. Стрелка указывает направление распространения ошибки. Развертка сети по времени производится на 2 шага.

Для обучения рекуррентной модели используется алгоритм распространения ошибки обратно по времени (backpropagation through time, BPTT). Уравнения для алгоритма распространения ошибки обратно по времени получаются в явном виде при использовании градиентного спуска. В конечной формулировке алгоритм фактически представляет собой алгоритм обратного распространения ошибки для сети «развернутой» во времени (рис.2).

Данный алгоритм обладает рядом ограничений. Доказано, что данная архитектура подвержена проблеме *затухания* или *всплеска градиента*. В зависимости от максимального собственного значения матрицы весов скрытого слоя W, норма градиента активации скрытого слоя в момент t как функции от активации в момент k либо экспоненциально растет, либо экспоненциально затухает с ростом $t - k$:

$$\frac{\partial h_t}{\partial h_k} = \prod_{k<i\leq t} W^T \cdot diag(f'(h_{i-1})) \tag{1}$$

В оригинальных работах по языковому моделированию при помощи рекуррентных нейронных сетей не предпринималось попыток каким-либо образом уменьшить эффект затухания градиента. Предполагалось, что в алгоритме распространения ошибки обратно по времени оптимальной глубиной развертки является 6 шагов, после чего прирост качества не наблюдался. Однако достаточно легко найти пример фразы, успешное предсказание слов которой может быть осуществлено лишь с учетом более длинных зависимостей:

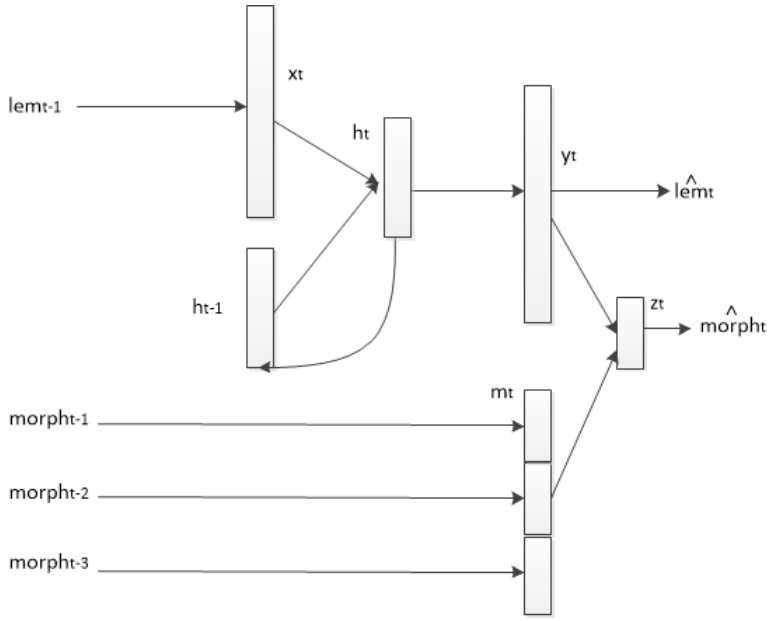

Рисунок 3: Рекуррентная нейронная сеть с внешним классификатором. m_t — конкатенация бинарных векторов с ненулевой координатой, соответствующей номеру вхождения грамматической пометки в списке допустимых пометок (словаре) GL; z_t — вектор длины $|GL|$ — оценка распределения грамматических форм для слова w_t

Студент московского ордена Ленина, ордена Октябрьской революции и ордена Трудового Красного знамени государственного университета им. М. В. Ломоносова получил красный диплом.

Данная проблема особенно актуальна для русского языка ввиду необходимости учитывать согласование слов в предложении.

При наличии словаря существенного объема статистическое моделирование флективных языков составляет дополнительную техническую проблему для нейросетевого подхода. Большое количество различных словоформ приводит к пропорционально большему размеру выходного слоя, причем сложность алгоритма обучения линейна по объему выходного слоя.

Чтобы обойти эту проблему, можно было бы использовать схему на рис.3. Каждое входное слово предварительно лемматизуется внешним морфологическим анализатором. Леммы используются для предсказания последующих лемм. Далее для предсказанной леммы запускается линейный классификатор (например, логистическая регрессия), предсказывающий словоформу по лемме и морфологическим признакам контекста. Данный подход позволяет миновать проблему разрастания словаря. Другой состоит в том, чтобы разделить выходной слой на два вектора — словарный (леммы) и морфологический (морфологические признаки). Ошибка предсказания в данном случае получается суммированием ошибок на двух векторах.

Основным результатом второй главы является экспериментальное подтверждение гипотезы о том, что языковая модель на рекуррентной нейронной сети для моделирования русского языка без учета морфологии оказывается более эффективной в смысле перплексии, чем 5-граммная модель со сглаживанием Кнесера-Нея на аналогичном лемматизованном корпусе. Эксперименты по перплексии были поставлены на лемматизованном корпусе новостных

заметок Lenta.ru объемом около 2 млн. словоупотреблений. Гипотезы распознавания в эксперименте по ранжированию были сгенерированы закрытой коммерческой системой распознавания речи на русском языке. Эксперимент демонстрирует, что проблема свободного порядка слов не является существенной для данной модели. Рекуррентная модель оказывается более эффективной, чем 5-граммная модель со сглаживанием Кнесера-Нея. Преимущество наблюдается как в эксперименте по измерению перплексии, так и в эксперименте по ранжированию гипотез распознавания.

Таким образом, рекуррентные нейронные сети являются перспективным инструментом для моделирования флективных языков. Тем не менее, исходная архитектура должна быть модифицирована для обеспечения поддержки большого словаря и учета длинного контекста.

В **третьей главе** предлагается оригинальная языковая модель, основанная на признаках, полученных из рекуррентной нейронной сети и тематического разложения левого контекста текущего слова.

Различные подходы к векторному представлению слов и контекстов приобрели популярность среди исследователей в связи с появлением модели word2vec и ее программной реализацией, выпущенной компанией Google.

Другим способом получения векторного представления контекста является вероятностное тематическое моделирование (PLSA). Данный подход основан на разложении матрицы частот слов в документах (матрица «термин-документ») $W = \Phi\Theta$ с дополнительным ограничением: матрицы Φ и Θ должны быть стохастическими, а их столбцы должны представлять дискретные вероятностные распределения $\theta_{t,d} = p(t|d)$, $\phi_{w,t} = p(w|t)$. Преимуществом вероятностного тематического моделирования является возможность учета различных факторов и наложения на модель специфических ограничений в зависимости от решаемой задачи.

Поиск матриц Φ и Θ сводится к минимизации расстояния Кульбака-Лейблера между эмпирическими оценками вероятностей $\hat{p}(w|d) = \frac{C(w,d)}{\sum_w C(w,d)}$ и вероятностью в текущей тематической модели $p(w,d) = \sum_{t \in T} p(w|t)p(t|d)$:

$$\sum_{d \in D} C(d) KL(\frac{C(w,d)}{C(d)} || \phi_{w,t}\theta_{t,d}) \to min,$$

где $C(w,d)$ — количество вхождений слова w в документ d, $C(d) = \sum_w C(w,d)$.

Обучение модели производится с помощью EM-алгоритма. EM-алгоритм также используется при разложении нового документа с помощью существующей модели. Осуществлять разложение нового документа необходимо, поскольку по мере обработки входного текста статистическая языковая модель должна вычислять новые вложения левого контекста в пространство тем, т.е. осуществлять тематическое разложение левого контекста.

Неотрицательное (стохастическое) матричное разложение $\Phi\Theta$ не является единственным и определено с точностью до невырожденного преобразования:

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta)$$

при условии, что матрицы $\Phi' = \Phi S$ и $\Theta' = S^{-1}\Theta$ также стохастические.

Таким образом, задача является некорректно поставленной. Общий подход к решению некорректно поставленных задач заключается во введении некоторых дополнительных ограничений —регуляризаторов — на параметры Φ, Θ, сужая тем самым множество решений. Использование аддитивной регуляризации позволяет перейти к задаче многокритериальной оптимизации и таким образом получить модель, обладающую дополнительными свойствами — разреженностью распределений ϕ_i и θ_j, контрастностью тем, а также устойчивостью модели к выбросам.

Основными регуляризаторами, используемыми для разработки предлагаемой языковой модели, являются сглаживающий, разреживающий и декоррелирующий регуляризаторы.

Сглаживающий регуляризатор минимизирует дивергенцию Кульбака-Лейблера между неизвестным распределением ϕ_{wt} или θ_{td} и некоторым априорным распределением β_w или α_t.

Согласно гипотезе разреженности, каждый документ и каждый термин принадлежат небольшому числу тем. С практической точки зрения также предпочтительными являются модели с сильно разреженными матрицами Φ и Θ, в которых доля нулевых значений превышает 90%. Разреживания можно добиться использованием энтропийного регуляризатора, максимизирующего KL-дивергенцию между равномерными распределениями β_w и α_t и распределениями ϕ_t и θ_d.

Повышение смыслового различия тем обеспечивается использованием декоррелирующего регуляризатора, минимизирующего ковариацию распределений ϕ.

Для повышения устойчивости тематической модели некоторые из тестируемых модификаций были расширены за счет шумовой и фоновой тем, определяемых соответственно как тема, содержащая слова общей лексики, и тема, содержащая редко встречающиеся слова. Выделение этих тем производилось автоматически с помощью аддитивной регуляризации.

Для выделения фоновой темы использовались: 1) сглаживающий регуляризатор, приближающий распределение слов в фоновой теме к распределению слов в коллекции β_w; 2) сглаживающий регуляризатор, приближающий распределение фоновой темы по документам к равномерому; 3) декоррелирующий регуляризатор, понижающий вероятности предметных тематических слов в фоновой теме.

Для получения шумовой темы использовались: 1) сглаживающий регуляризатор, приближающий распределение слов в фоновой теме к некоторому распределению β'_w, в котором вероятность слова w *обратно* пропорциональна частоте по коллекции; 2)сглаживающий регуляризатор, приближающий распределение шумовой темы по документам к равномерому.

Предлагаемая статистическая языковая модель (рис.4) предполагает отдельное обучение рекуррентной модели с последующим объединением признаков, полученных из скрытого слоя сети на шаге t с тематическим профилем контекста на шаге t в рамках модели максимальной энтропии.

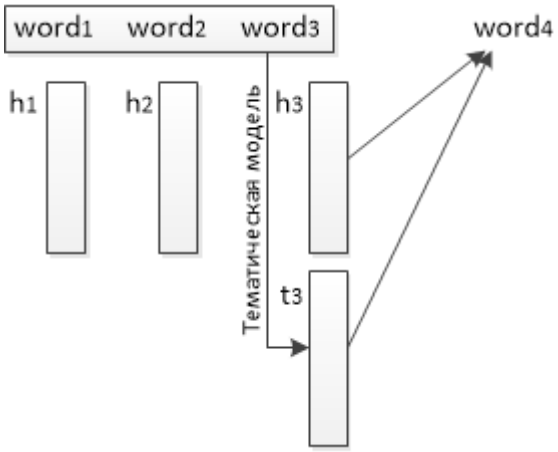

Рисунок 4: Схема работы предлагаемой модели. Рекуррентная нейронная сеть генерирует последовательность векторов скрытых состояний h_t, тематическая модель генерирует последовательность тематических разложений левого контекста t_3

$$p(w_{t+1}|h_t, d_t) = \frac{e^{-v_w \cdot h_t - f_w \cdot d_t}}{\sum_{w'} e^{-v_{w'} \cdot h_t + f_{w'} \cdot d_t}},$$

где v_w — вектор весов слова w для элементов вектора скрытого состояния h_t, f_w — вектора весов слова w для вектора тематического разложения левого контекста d_t.

Как и в случае с нейронной сетью, процедура обучения модели максимальной энтропии может быть весьма длительной, поскольку задействует большое количество матричных операций. По этой причине как рекуррентная нейронная сеть, так и модель максимальной энтропии были реализованы на GPU.

Тематическая модель была обучена на новостном корпусе Lenta.ru (апрель 2014 — март 2015).

Для экспериментов было обучено три тематических модели:

1. Модель с комбинацией сглаживающих, разреживающих и декоррелирующего регуляризаторов и дополнительными предположениями о шумовой и фоновой темах (srPLSA);
2. Модель из предыдущего пункта, но без шумовой и фоновой тем (sPLSA);
3. Модель только со сглаживающим регуляризатором (аналог LDA).

Модель на рекуррентной нейронной сети и модель максимальной энтропии были натренированы на подкорпусе корпуса Lenta.ru объемом приблизительно в $2 \cdot 10^6$ токенов. Примерно 10% данных было выделено для валидации. Каждый текст был обработан морфологическим анализатором/лемматизатором для русского языка со встроенным словарем примерно в $2{\cdot}10^6$ словоформ. Выходом анализатора являлся текст, в котором все известные словоформы были заменены соответствующими леммами, а неизвестные — специальным токеном "UNK". Для получения разложения левого контекста использовалось скользящее контекстное окно. Тестировались различные варианты шага вычисления разложения и длины контекстного окна. Размер скрытого слоя был установлен равным 500.

Для каждой из моделей измерялась перплексия на тестовой выборке. Результаты в виде зависимости перплексии данных от длины окна конеткста показаны на рис.5.

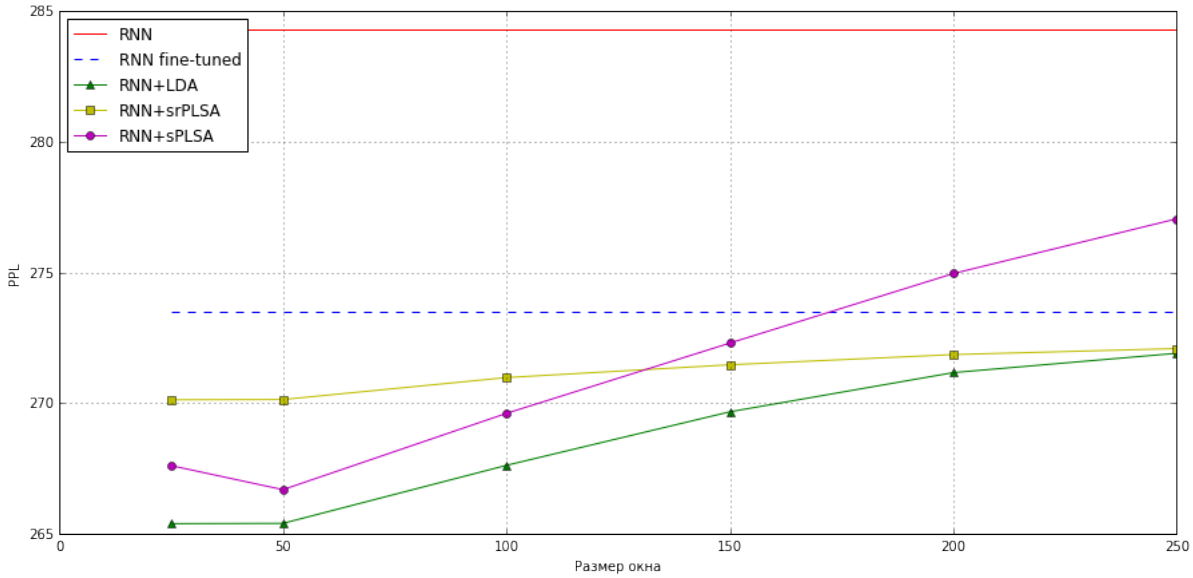

Рисунок 5: Зависимость качества языковой модели на основе тематического разложения от длины контекстного окна. Во всех случаях вложения вычислены с шагом $F_e = 25$.

После уменьшения шага вложения с 25 до 5 наилучший результат, достигнутый в экспериментах с длиной окна $L = 25$, был улучшен на 4.2%. Таким образом, перплексия достигла отметки 254.23 . Улучшение относительно исходной модели составило 10.56%, относительно дообученной модели — соответственно 7.04%. Можно предложить как минимум две причины улучшения качества относительно модели с большим шагом вложения. Первая из них состоит в том, что тема, действительно, меняется достаточно быстро и вычисление необходимо производить чаще. Вторая причина может состоять в том, что с уменьшением шага вложения увеличивается количество различных тематических векторов в обучающей выборке, что приводит к лучшей обобщающей способности модели.

Наиболее эффективными для языкового моделирования являются признаки, полученные на основе модели LDA. Тем не менее, вероятностная тематическая модель с разреживанием позволяет добиться почти такого же уровня перплексии. Использование резреженности может быть важным фактором для хранения моделей большого размера.

Четвертая глава посвящена предсказанию морфологических характеристик леммы. Поскольку предсказание морфологических характеристик является более простой задачей в смысле перплексии, а зависимости между элементами цепочки более короткими, использование рекуррентной архитектуры на данном этапе не является необходимым. В то же время более сильная зависимость между соседними элементами в последовательности морфологических пометок указывает на целесообразность использования сверточных слоев.

В главе приводятся данные эксперимента с рекуррентной сетью, осуществляющей одновременное предсказание лемм и морфологических характеристик. Эксперимент показал, что обучение данной архитектуры затруднено, и результатов, сопоставимых с показателями пер-

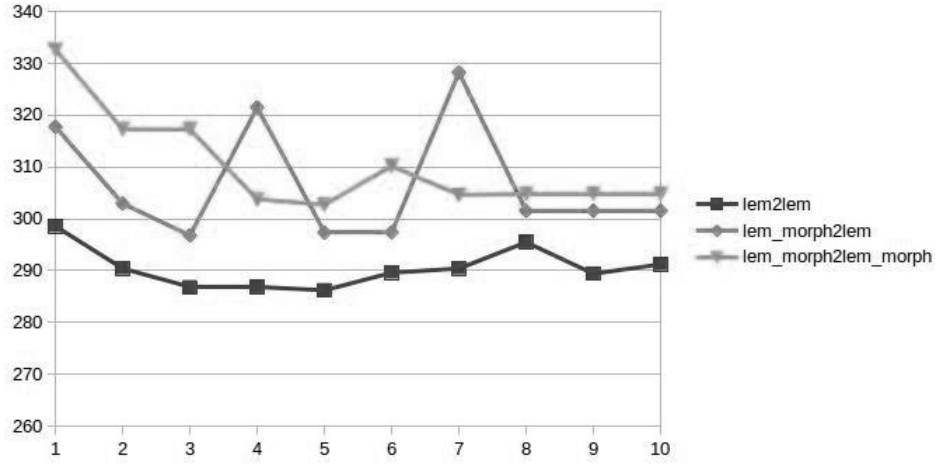

Рисунок 6: Зависимость перплексии по леммам от размера скрытого слоя для различных архитектур. Одно деление на оси абсцисс соответствует 100 элементам на скрытом слое

плексии для сглаженных n-граммных моделей, достичь не удалось. Более простая модель с входным слоем, содержащим морфологические характеристики, но осуществляющая предсказание только лемм, также дала худший результат в сравнении с моделью без морфологии. Как и в предыдущих главах, для эксперимента использовался корпус новостей Lenta.ru.

Принимая во внимание этот результат, можно сформулировать дальнейшую задачу как предсказание морфологических признаков леммы по известному левому контексту: $\mathsf{P}(m_t|l_1 \dots l_t, m_1 \dots m_{t-1})$, где лемма l_t предсказывается рекуррентной нейронной сетью, описанной в главе 2.

Важным отличием задачи определения морфологических признаков от предсказания леммы является большое различие в количестве классов (15000 для лемм и не более 200 для морфологических признаков) и характер зависимостей в последовательности: зависимость между морфологическими признаками слов в предложении действует на более коротких расстояниях. Это позволяет отказаться от использования рекуррентной архитектуры, однако приводит к необходимости отбора и конструирования сложных признаков. Дизайн признаков в свою очередь является достаточно трудоемкой задачей, требующей принятия априорных предположений о том, какие из морфологических характеристик могут влиять друг на друга.

Использование глубоких нейронных сетей позволяет избежать ручного дизайна признаков. Это оправдывает использование глубокого обучения для задачи предсказания морфологических характеристик.

Подход к решению задачи предсказания морфологических характеристик на основе сверточных нейронных сетей описан ниже.

Рассмотрим цепочку слов $w_1 \dots w_N$. Обозначим за $v(w_i)$ d-мерный вектор, однозначно соответствующий данному слову. Метод отображения $\mathbb{V} \to \mathbb{R}^d$ вообще говоря не важен. N-

грамме $w_1 \dots w_N$ можно поставить в соответствие матрицу размера $N \times d$:

$$S = \begin{bmatrix} — & v(w_1)^T & — \\ — & v(w_2)^T & — \\ & \vdots & \\ — & v(w_N)^T & — \end{bmatrix}$$

Тогда двумерная свертка с коэффициентами фильтра размера $r \times d$ и последующим применением нелинейной функции активации f задает для каждой r-граммы последовательности $w_1 \dots w_N$ отклик на эту r-грамму. Задав K фильтров, получим отображение всей последовательности в новое признаковое пространство согласно формуле:

$$c_{k,i} = f(\sum_{r'=1}^{r} \sum_{j=1}^{d} a_{rjk} v(w_{i+r'-1})_j),$$

где a_{rjk} — параметры модели, f — нелинейная функция.

Данный подход является стандартным при обработке изображений в сверточных нейронных сетях.

Модель предсказания морфологии с помощью сверточной нейронной сети опирается на архитектуру нейронной сети с одним сверточным и несколькими полносвязными слоями. Последний слой осуществляет классификацию на N классов, где N — количество возможных морфологических тегов для данной части речи.

Каждый элемент n-граммы на шаге t можно представить в виде вектора $m(k) \in \{0,1\}^{|G|}$, $t - n + 2 < k < t + 1$, где $\mathbb{G}$ — множество грамматических помет. Элемент $m_i(k) = 1$ тогда и только тогда, когда тег на шаге t содержит i-ю морфологическую метку.

Добавление семантических признаков обеспечивается за счет расширения вектора $m(k)$ двумя векторами $v(k)$ и $v(t+1)$, полученными в отдельно обученной word2vec-модели для русских словоформ (корпус LibRuSec). $v(k)$ представляет собой векторное представление леммы на шаге k, вектор $v(t + 1)$ — соответственно, векторное представление известной леммы на шаге $t + 1$, морфологический тег которой необходимо предсказать. Таким образом, размерность представления каждого элемента в n-грамме составляет $2d + |G|$, где d — размерность векторного представления лемм.

Входом классификатора является матрица $S_{n\times(2d+|G|)}$, причем $\forall i S_{i,1:d} = v(t + 1)$. Данное решение позволяет отразить в сверточной архитектуре факт взаимосвязи каждого элемента n-граммы с целевым словом.

Поскольку лемма и часть речи предполагаются известными, можно обучить отдельный классификатор для каждой части речи, что позволит полностью исключить возможность предсказания невозможных тегов для входной части речи. Очевидно, что в этой ситуации размер выходного слоя и количество классов различны, что сказывается на значении кросс-энтропии.

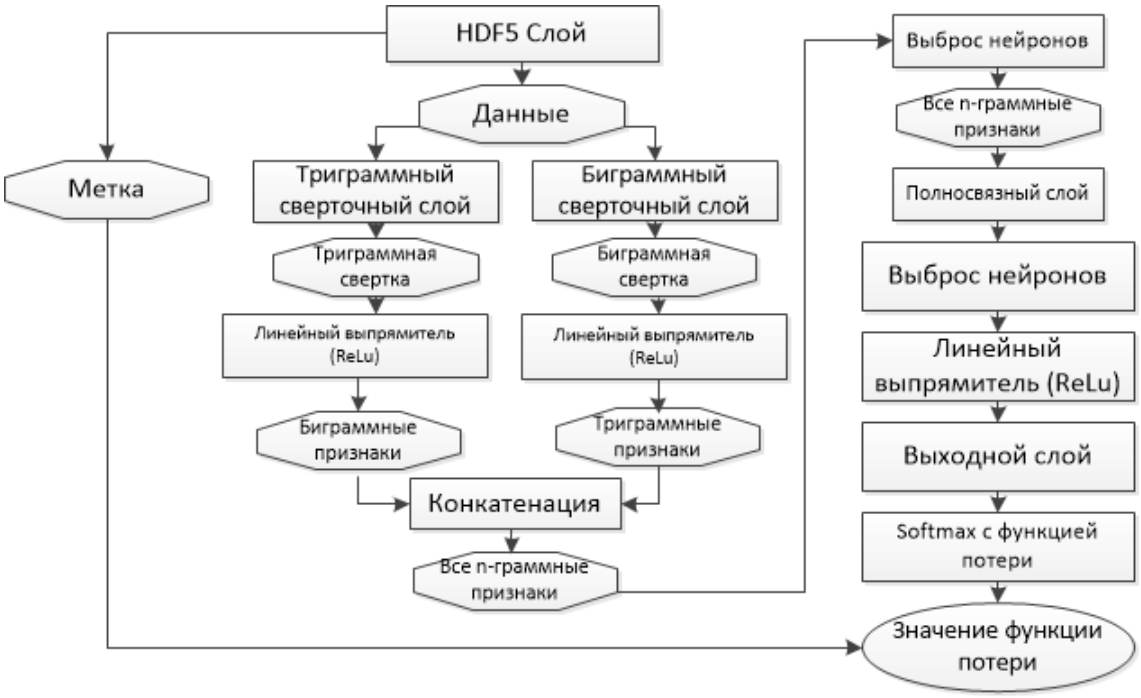

Рисунок 7: Схема сверточной нейронной сети для предсказания морфологии в библиотеке Caffe

Таблица 1: Кросс-энтропия на тестовой выборке при предсказании точных форм для изменяемых частей речи

Часть речи (пометка)	**Количество классов**	**Кросс-энтропия**
Существительное (S)	113	0.718
Глагол (V)	118	0.938
Наречие (ADV)	3	0.037
Прилагательное (A)	32	1.067
Числительное (NUM)	48	1.025

Эксперименты со сверточными сетями проводились при помощи пакета Caffe. Схема слоев сети в программе Caffe приведена на рис.7.

Как и в предыдущих экспериментах, для обучения и тестирования использовался новостной корпус сайта Lenta.ru за 2014 год. Объем корпуса составил 3211256 токенов. Корпус был обработан морфологическим анализатором и состоял из последовательностей пар вида «лемма:тег».

Результаты тренировки морфологических классификаторов для изменяемых частей речи приведены в таблице 1. Стоит отметить, что на момент остановки процедуры обучения снижение перплексии на валидационной выборке все еще продолжалось, и в дальнейших экспериментах результат может быть улучшен.

Для проверки результатов модели в эксперименте по распознаванию речи использовалась свободная библиотека Kaldi. Kaldi допускает использование переранжирования гипотез, полученных с помощью первичной n-граммной модели.

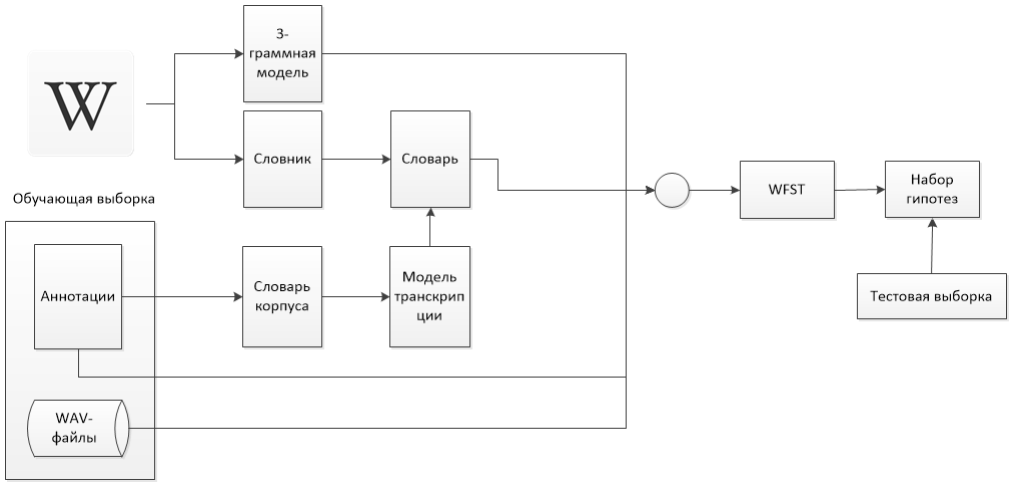

Рисунок 8: Схема процесса подготовки данных

Для экспериментов по распознаванию речи использовался речевой корпус на русском языке со следующим составом дикторов: в том числе 127 мужчин и 110 женщин в возрасте от 18 до 65 лет. Каждый диктор произносил по 70 предложений.

В качестве n-грамной модели для Kaldi была выбрана 3-граммная модель со сглаживанием Кнесера-Нея, обученная на подмножестве статей архива русского раздела сайта «Википедия». Объем обучающего корпуса составил около 20 млн. словоформ. Тренировка модели осуществлялась с помощью пакета SRILM. Объем словника составил порядка 100 тыс. словоформ.

Для получения транскрибированного словаря из словника была использована специально написанная утилита для расстановки ударений и автоматический транскриптор, основанный на скрытых марковских моделях.

Общая схема подготовки данных приведена на рис. 8.

Наилучшая базовая модель с трифонами, обученная с помощью Kaldi, показала WER 17.8%.

Результаты эксперимента по переранжрованию приведены в таблице 2.

В поставленном эксперименте рекуррентная модель на леммах не дала практически никакого улучшения. Данный факт можно объяснить значительным различием размеров обучающих выборок в случае 3-граммной и рекуррентной моделей: размер обучающего корпуса отличался в 10 раз. Несмотря на то, что для тренировки модели использовалась реализация на GPU, для обучения модели на корпусе сопоставимого размера по-прежнему требуется время на порядок превосходящее время тренировки сглаженной n-граммной модели. Необходимо отметить, что в эксперименте не были задействованы более поздние модификации модели, направленные на ускорение обучения, так как такие модификации приводят к снижению качества по сравнению с канонической моделью.

Использование сверточной модели для предсказания морфологии приводит к улучшению качества распознавания, при этом есть основания считать что с дообучением модели результат может быть улучшен. Тем не менее, окончательный результат может быть получен только

Таблица 2: Результаты эксперимента по переранжированию гипотез. CNN — сверточная нейронная сеть для предсказания морфологии. λ — коэффициент при рекуррентной и сверточной моделях в интерполяции, приводящий к наибольшему падению WER

Модель	λ	**WER**, %
3-gram	0	17.80
3-gram + CNN	0.2	17.2

при наличии рекуррентной модели для предсказания лексики с размером словаря и объемом обучающей выборки, соответствующим размеру словаря и тематическому разнообразию речевого корпуса. Данный эксперимент будет поставлен в будущем. В случае получения положительного результата в эксперименте с использованием рекуррентной модели на леммах, станет возможным постановка эксперимента с моделью, использующей тематическое моделирование.

Заключение

В работе было выполнено исследование возможностей построения вычислительно-эффективной статистической модели русского языка с высокой точностью предсказаний с целью использования в системах распознавания речи и текстового ввода на мобильном устройстве. В соответствии с поставленной задачей в диссертационной работе получены следующие результаты:

1. Предложена языковая модель, предполагающая раздельное предсказание лемм и морфологии на основе нейронных сетей с целью уменьшения количества классов, соответствующих словоформам.

2. Предложена и протестирована модель предсказания лемм на основе рекуррентной нейронной сети для русского языка.

3. Впервые предложена и проанализирована гибридная языковая модель максимальной энтропии, основанная на рекуррентной нейронной сети и вероятностном тематическом моделировании.

4. Впервые предложена и проанализирована модель классификатора для предсказания морфологии с помощью сверточной нейронной сети.

Результаты указывают на перспективность дальнейших исследований в данной области, включая эксперименты с другими рекуррентными моделями и модификациями исходной модели, направленными на улучшение быстродействия.

Результаты работы могут быть использованы в программных продуктах, задействующих языковые модели, таких как распознавание речи, распознавание текста, проверка орфографии и предиктивный ввод.

Публикации автора во теме диссертации

Статьи в журналах, рекомендованных ВАК

1. *Kudinov M., Romanenko A., Piontkovskaya I.* Conditional Random Fields in segmentation and noun phrase inclincation tasks for Russian // Computational linguistics and intellectual technologies: Proceedings of the International Conference Dialogue — 2014. — Pp. 297–307.

2. *Kudinov M., Piontkovskaya I.* Automatic update of the named entities database based on users queries // Computational linguistics and intellectual technologies: Proceedings of the International Conference Dialogue — 2015. — Vol.1 — Pp. 369–376.

3. *Kudinov M.* Recurrent Neural Networks for Hypotheses Re-Scoring // Speech and Computer — 17th International Conference, SPECOM 2015, Athens, Greece, September 20-24, 2015, — 2015. — Pp. 341–347.

4. *Kudinov M., Romanenko A.* Hybrid language model based on recurrent neural network and probabilistic topic modeling // Pattern Recognition and Image Analysis. — 2016. — Vol. 26. — Pp.587–592.

5. *Кудинов М. С.* Использование рекуррентных нейронных сетей для ранжирования списка гипотез в системах распознавания речи // Современная наука: актуальные проблемы теории и практики. Серия «Естественные и технические науки. — 2016. — С.358–364.

Статьи в прочих журналах

6. *Кудинов М.* Частичный синтаксический разбор текста на русском языке с помощью условных случайных полей // Машинное обучение и анализ данных. — 2013. — Т.1., №6 — С.714–724

Тезисы в докладах на конференциях

7. *Кудинов М.* Частичный синтаксический разбор текста на русском языке с помощью условных случайных полей // Математические методы распознавания образов: XVI Всероссийская конференция, г.Казань 6–12 сентября 2013 г.: Тезисы докладов — М.: Торус Пресс, 2013. — С.93