Journal of Siberian Federal University. Engineering & Technologies, 2016, 9(8), 1291-1301
~ ~ ~

УДК 004.934.2

On Applicability of Recurrent Neural Networks to Language Modelling for Inflective Languages

Mikhail S. Kudinov*
Samsung R&D Center Russia
1 corp., 12, Dvintsev, Moscow, 127018, Russia

Received 20.04.2016, received in revised form 09.06.2016, accepted 12.09.2016

Standard version of recurrent neural network language model (RNNLM) has shown modest results in language modelling of Russian. In this paper we present a special modification of RNNLM making separate predictions of lemmas and morphology. New model shows superior results compared to Knesser-Ney language model both in perplexity and in ranking experiment. At the same time morphology integration has not shown any improvement.

Keywords: language models, recurrent neural network, inflected languages, speech recognition.

Citation: Kudinov M.S. On applicability of recurrent neural networks to language modelling for inflective languages, J. Sib. Fed. Univ. Eng. technol., 2016, 9(8), 1291-1301. DOI: 10.17516/1999-494X-2016-9-8-1291-1301.

О применимости рекуррентных нейронных сетей к задаче статистического моделирования русского языка

М.С. Кудинов
ООО «Исследовательский центр Samsung»
Россия, 127018, Москва, ул. Двинцев, 12, стр. 1

В статье представлены данные экспериментов по использованию рекуррентных нейронных сетей для языкового моделирования русского языка. Ранее уже была продемонстрирована невысокая эффективность стандартной архитектуры рекуррентной нейронной сети для моделирования русского языка. В данной статье рассматривается модель, осуществляющая предсказание леммы и морфологии последующего слова отдельно. Показано, что модель, использующая только леммы, превосходит n-граммную модель Кнессера-Нея как по перплексии, так и в простом эксперименте по ранжированию гипотез в распознавании речи. В то же время попытки внедрения морфологии в обучение нейронной сети не приводят к улучшениям.

Ключевые слова: языковые модели, рекуррентная нейронная сеть, флективные языки, распознавание речи.

* Corresponding author E-mail address: mikhailkudinov@gmail.com

Введение

Известно, что проблема статистического моделирования флективных языков представляет большую сложность, чем английского языка [1]. Основные проблемы возникают вследствие большого количества морфологических форм слов (лемм) и более свободного порядка слов [2]. Обе проблемы в результате усиливают разреженность данных и снижают эффективность *n*-граммных моделей.

В то время как использование *n*-граммных моделей на первых стадиях распознавания сегодня является стандартной практикой [3], возможности для последующей обработки в рамках алгоритма распознавания, осуществляющего несколько проходов по входным данным, гораздо шире. Например, для переранжирования гипотез, возвращаемых процедурой лучевого поиска Витерби, может быть использована морфологическая, синтаксическая и семантическая информация. В последнем случае значения слов представляются посредством вложения слов в некоторое векторное пространство. К методам, осуществляющим такие вложения, относятся: латентно-семантический анализ [4], вероятностное тематическое моделирование [5] или нейронные сети [6]. В 2010 г. была представлена языковая модель на рекуррентной нейронной сети (RNNLM) [7]. Использование данной модели позволило улучшить предыдущие результаты на стандартных наборах данных как в перплексии, так и в пословной ошибке в экспериментах по распознаванию речи. Несмотря на то что модель была предложена для английского языка, в [8] были приведены обнадеживающие результаты, полученные на небольшом наборе данных для чешского языка. Сходство чешского и русского языков общеизвестно, а значит, перспективы применения рекуррентных нейронных сетей к русскому материалу выглядят многообещающе. Тем не менее эксперименты в [9] продемонстрировали в целом невысокую эффективность данной модели для русского языка. Параметры, используемые авторами, впрочем, не выглядят оптимальными с точки зрения качества модели, однако выбор именно таких параметров был, очевидно, продиктован необходимостью поддержки большого словаря – списка потенциальных словоформ.

Таким образом, проблема обучения рекуррентной нейронной сети для языков с богатой морфологией является более сложной, по крайней мере, если использовать оригинальный подход из [7]. В дополнение к уже упомянутым трудностям, связанным с разреженностью данных, обучение модели, применяющий словник, содержащий все допустимые словоформы, потребовало бы слишком длительного времени. Более перспективным в этой связи выглядит использование сложных векторных моделей, отражающих сходство семантики слов [10, 11], для предсказания лемм с последующим выбором морфологической формы на основании более простых моделей. В данной статье рассматривается несколько рекуррентных архитектур, различающихся использованием морфологической информации: рассмотрена модель, полностью игнорирующая морфологию; модель, использующая морфологические признаки для предсказания лемм, и ее модификация, дополнительно осуществляющая предсказание морфологической формы.

Методология исследования

Рекуррентные нейронные сети впервые были рассмотрены Элманом в 1990 г. [12]. В данном исследовании также была высказана идея о применимости рекуррентной нейронной сети

для моделирования языка. Тем не менее вследствие значительной вычислительной сложности и отсутствия доступных лингвистических корпусов достаточного объема на тот момент метод не получил широкого распространения.

Другой важной вехой в развитии нейросетевых языковых моделей является работа И. Бенджио (2003), в которой предлагается метод предсказания последующего слова по левому контексту длины n-1, таким образом формируя своего рода n-граммную нейросетевую модель n-го порядка. Однако в отличие от n-граммной модели в данном случае предсказание осуществляется на основании вложений слов в векторное пространство R^M. Каждое входное слово (допустим, с индексом l) в словаре объемом $|L|$ слов представляется в виде $|L|$-мерного вектора $w =< 0_1,...,1_l,0_{l+1},...,0_{|L|} >$ с единственной ненулевой координатой $w_l = 1$. На вектор слева умножается матрица U *размерности* $M \times |L|$, что эквивалентно выборке l-го столбца U. Другими словами, U действует как словарная таблица, осуществляющая однозначное отображение слов на их векторные представления.

Аналогичная техника была применена Т. Миколовым, который использовал рекуррентную сеть Элмана для предсказания слов по контексту [7]. Результирующая модель описывалась следующими уравнениями:

$$P(w_k|w_{t-1}, h_{t-1}) = y_{w_k}(t), \tag{1}$$

$$y(t) = s(V \cdot h_t), \tag{2}$$

$$h_t = \sigma(U \cdot x + W \cdot h_{t-1}), \tag{3}$$

где

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

– логистическая функция активации, а

$$s(x) = \frac{e^x}{\sum_i e^{x_i}}$$

– софтмакс-функция, x_t – вектор с единственной единичной координатой; h_t – рекуррентный слой; y – выходной слой, где каждому k-му элементу соответствует вероятность $P(w_k|w_{t-1},h_{t-1})$ ($W_{H\times H}$ – матрица весов рекуррентного слоя, $U_{H\times|L|}$ – словарная таблица, отображающая слова в векторные представления, $V_{|L|\times H}$ – матрица весов выходного слоя); H – количество нейронов скрытого слоя (рис. 1).

Поскольку h_t потенциально сохраняет в себе весь левый контекст, данная модель выглядит более мощной, чем n-граммная нейросетевая модель. К сожалению, в действительности последнее утверждение не совсем верно, поскольку норма градиента $\frac{\partial h_t}{\partial h_k}$, $k<t$, отражающего влияние предыдущих значений на скрытом слое на последующие, стремится к нулю (или к бесконечности) с экспоненциальной скоростью по $(t - k)$:

$$\frac{\partial h_i}{\partial h_k} = \prod_{k<i\leq t} W^T diag\left(\sigma'(h_{i-1})\right), \tag{4}$$

где $diag(f(x))$ обозначает диагональную матрицу с элементами на главной диагонали, вычисляемыми по формуле $A_{i,i} = f(x_i)$ [13, 14].

В зависимости от свойств матрицы W значение выражения (4) либо растет, либо падает с экспоненциальной скоростью. Данный факт получил название затухания градиента (*vanishing gradient*) в случае убывания или градиентного взрыва (*gradient explosion*) в случае роста [10, 12].

Фактически данный результат означает, что тренировка рекуррентной нейронной сети методами первого порядка не может учитывать влияния элементов последовательности, если они сильно разнесены по времени. Для этого матрица W должна была бы иметь достаточно большую норму, а значит, быть критически восприимчивой к шуму в обучающей последовательности [10]. На практике это выражается в высокой амплитуде норм градиентов и неустойчивости решения. С другой стороны, устойчивое решение может быть получено при небольших нормах W, однако, как было показано выше, такие решения приводят к сложностям с моделированием дальних зависимостей.

Хотя за прошедшие 20 лет с момента обоснования данной проблемы было предложено немало способов ее решения [14, 15], в [7] утверждается, что данная проблема не является существенной для моделирования языка. Таким образом, в данной cnfnmt будет рассмотрен случай стандартной архитектуры Элмана с алгоритмом распространения ошибки обратно по времени (backpropagation through time).

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, а из (3) видно, что сложность алгоритма обучения линейна по объему выходного слоя.

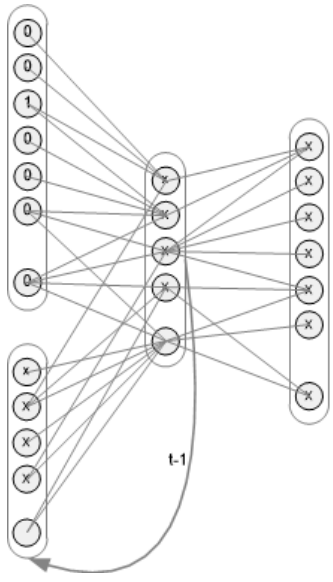

Рис. 1. Рекуррентная нейронная сеть для статистического моделирования языка

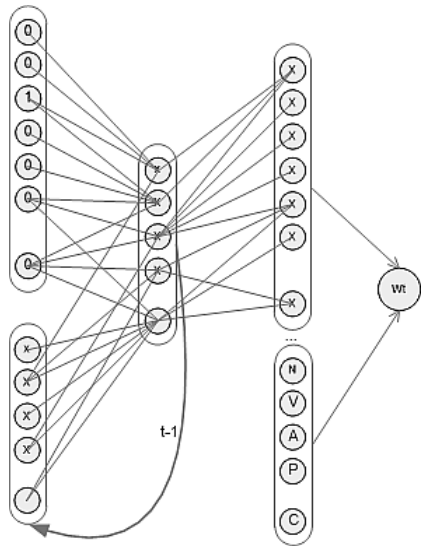

Рис. 2. Рекурентная нейронная сеть с внешним классификатором

Эксперименты [9] показали, что для достижения достаточного покрытия словаря необходимо использовать неоптимальные с точки зрения перплексии параметры обучения, которые позволяют провести обучение за разумное время. Кроме того, авторы используют готовую утилиту Т. Миколова, реализующую дорогостоящие матричные вычисления в один поток, что еще больше замедляет обучение. Тем не менее данный результат можно считать экспериментальным подтверждением того, что прямое применение рекуррентной нейронной сети для флективных языков затруднено.

Чтобы обойти эту проблему, можно было бы использовать схему, представленную на рис. 2. Каждое входное слово предварительно лемматизуется внешним морфологическим анализатором. Леммы используются для предсказания последующих лемм. Далее предсказанной леммы запускается линейный классификатор (например, логистическая регрессия), предсказывающий словоформу по лемме и морфологическим признакам контекста. Данный подход позволяет миновать проблему разрастания словаря. Другой подход мог бы состоять в том, чтобы разделить выходной слой на два вектора – словарный (леммы) и морфологический (морфологические признаки). Ошибка предсказания в данном случае получалась бы суммированием ошибок на двух векторах.

Ниже будет показано, что как минимум второй из предложенных подходов не дает обнадеживающих результатов. Более того, даже простая модификация стандартной архитектуры с добавлением морфологических признаков приводит к росту перплексии. При этом рекуррентная нейросетевая модель, игнорирующая морфологию, т.е. модель, работающая на лемматизованном корпусе, работает лучше, чем *n*-граммная модель со сглаживанием Кнессера-Нея в аналогичных условиях.

Наконец, из простого эксперимента по ранжированию гипотез видно, что комбинация нейронных сетей, обученных на леммах, дает лучший результат, чем комбинация *n*-граммных моделей с дисконтированием Кнессера-Нея.

Эксперименты

В эксперименте рассматривались рекуррентные нейронные сети с тремя различными архитектурами: стандартная архитектура, игнорирующая морфологию (рис. 1), и архитектуры, использующие морфологию (рис. 3).

Архитектура первого типа представляет собой стандартную архитектуру Т. Миколова (здесь и далее l_t, m_t – соответственно лемма и список морфологических признаков словоформы виде *tag1@tag2@...@tagN* на шаге t): в данном эксперименте леммы левого контекста предсказывали последующие леммы: $P(l_t|l_{t-1},h_{t-1})$.

Архитектура второго типа отличается только наличием дополнительного вектора морфологических признаков. Таким образом, леммы и морфологические признаки левого контекста предсказывали последующие леммы: $P(l_t|l_{t-1},m_{t-1},h_{t-1})$.

Наконец, архитектура третьего типа осуществляет также предсказание морфологической формы: $P(l_t,m_t|l_{t-1},m_{t-1},h_{t-1})$.

Для эксперимента был подготовлен новостной корпус, основанный на заметках издания Lenta.ru за март-ноябрь 2014 г. Корпус насчитывал $1{,}8 \cdot 10^6$ словоупотреблений. Корпус был обработан морфологическим анализатором [16] и преобразован в формат последовательностей вида *лемма$_1$:морфология$_1$ лемма$_2$:морфология$_2$... лемма$_N$:морфология$_N$*. Словарь был ограничен 10 000 лемм. Остальные леммы получали метку «UNK». По 10 % корпуса были выделены для тестовой и валидационной выборок.

Каждая из архитектур обучалась на компьютере, снабженном CUDA-совместимой видеокартой NVIDIA GTX TITAN. Были протестированы различные объемы скрытого слоя

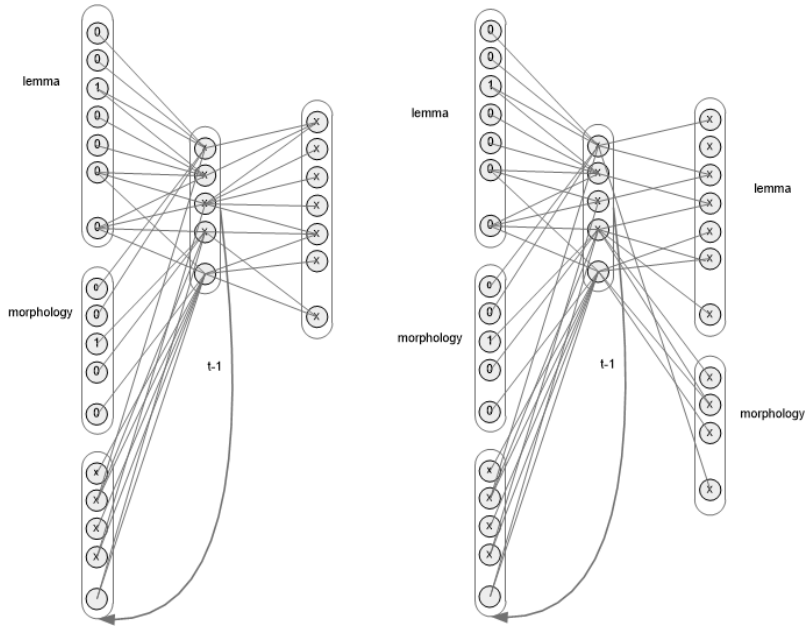

Рис. 3. Рекуррентные нейронные сети с морфологической информацией: слева – морфологические признаки используются для предсказания лемм; справа – морфологические признаки используются для предсказания лемм и морфологических форм

от 100 до 1000. Обучение каждой из сетей занимало около 20 ч почти вне зависимости от размера скрытого слоя, что является существенным улучшением по сравнению с используемой в [9] утилитой Т. Миколова [**17**], не применяющей параллельных матричных вычислений.

Ниже приведены результаты работы алгоритмов на валидационной выборке. Перплексии на тестовой и валидационной выборке обычно разнятся в пределах 3–4 пунктов.

Из табл. 1 и графиков на рис. 4 видно, что перплексии моделей, использующих морфологию, всегда выше, чем модели без морфологии. Если для модели номер 3 это можно было бы объяснить большей сложностью функции ошибки – фактически над одним и тем же множеством переменных строится два вероятностных распределения вместо одного, – то падение качества для второй модели выглядит неожиданным. Одной из причин понижения качества могло бы стать переобучение вследствие увеличения числа параметров, однако различие в числе параметров нельзя назвать значительным ввиду небольшого числа морфологических классов в сравнении с количеством лемм.

Таблица 1. Перплексии моделей на валидационной выборке*

Модель Скрытый слой	Леммы Леммы	Леммы + Морфология→ Леммы	Леммы + Морфология→ Леммы + Морфология
100	298,58	317,78	332,68 / 20,922
200	290,42	302,96	317,36 / 19,16
300	286,80	296,81	317,36 / 18,49
400	286,82	321,50	303,80 / 18,83
500	286,22	297,35	302,75 / 18,95
600	289,62	297,40	310,19 / 18,48
700	290,42	328,26	304,64 / 18,62
800	295,49	301,58	304,86 / 18,64
900	289,41	–	–
1000	291,23	–	–

* Для второй и третьей моделей эксперименты с объемом слоя больше 800 не проводились. Для модели 3 приведены перплексии для предсказаний леммы и морфологической формы.

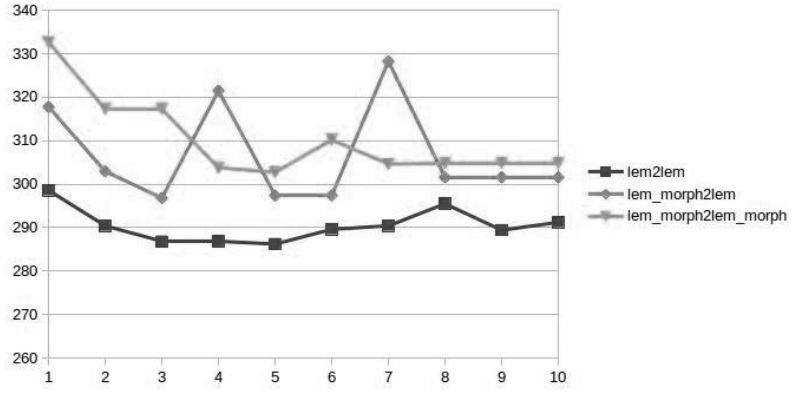

Рис. 4. Зависимость перплексии от размера скрытого слоя. lem2lem: Лемма→Лемма; lem_morph2lem: Лемма+Морфология→Лемма; lem_morph2lem_morph: Лемма+Морфология→Лемма+Морфология

Вполне вероятно, что более тщательный подбор гиперпараметров – количества нейронов на скрытом слое и параметров регуляризации – позволили бы получить значительно лучшее качество. Подбор гиперпараметров – крайне затратная по времени операция, поскольку количество итераций алгоритма обратного распространения ошибки, необходимых для оценки качества модели с данной конфигурацией гиперпараметров, как правило, довольно велико. При этом по результатам экспериментов нельзя выявить какой-либо тренд в зависимости перплексии на валидационной выборке от размера скрытого слоя. По крайней мере, нельзя с уверенностью утверждать, что увеличение размера скрытого слоя даст положительный эффект.

Таким образом, по результатам поставленного эксперимента можно сделать вывод, что добавление дополнительных признаков в рекуррентную нейронную сеть является самостоятельной задачей, требующей исследования. Для предсказания лемм модель, игнорирующая морфологию, в целом представляется более надежной. С другой стороны, необходимо найти решение для задачи предсказания морфологических характеристик.

Для второй серии экспериментов описанный выше корпус был повторно обработан для обучения *n*-граммной модели со сглаживанием Кнессера-Нея. Обработка заключалась в замене словоформ для лемм, не попадающих в словарь (10 000 лемм), на «UNK». Таким образом, в дополнение к корпусу для тренировки модели на леммах был получен корпус для словоформ. На полученных корпусах проводилось обучение и эксперименты по определению перплексии.

Для эксперимента по ранжированию гипотез применялись списки гипотез, полученные от внешней системы распознавания фирмы Nuance. Использовался русскоязычный корпус предложений со студийным качеством записи и транскрипциями. Аудиофайлы подавались на вход системе распознавания. На выходе получалось до 10 гипотез. В результате имеется коллекция неотсортированных списков гипотез. Как правило, список не содержал полностью правильной гипотезы и она добавлялась вручную.

Далее каждая гипотеза обрабатывалась теми же инструментами, которые использовались при подготовке корпусов, т.е. были проведены лемматизация и замены неизвестных слов. Полученные корпуса были обработаны обученными на предыдущем этапе моделями. В результате для каждой из гипотез были получены списки откликов от каждой модели – *n*-граммной со сглаживанием Кнессера-Нея и рекурентных нейронных сетей с различными размерами скрытого слоя. Всего в обучающем корпусе для ранжирования было 1300 фраз со средним значением 5 гипотез на фразу. В тестовом корпусе было 300 фраз.

В тестах были использованы *n*-граммные модели со сглаживанием Кнессера-Нея, порядков 3, 4, 5, натренированные на леммах и на словоформах. Модели на основе рекуррентных различались размером скрытого слоя. Были протестированы модели с объемами слоя 100, 200, 300, 400 и 500. Все рекуррентные сети обучались на лемматизованном корпусе. Кроме того, использовалась оценка, возвращаемая морфологическим анализатором. В результате было получено 12 оценок.

Для ранжирования брали модель ranking SVM, где в качестве признаков выступали оценки моделей. Результирующая модель обучалась ранжированию гипотез в списке на две категории — верная и неверная гипотеза. Фактически данный подход дает интерполяцию моделей. В качестве метрик для оценки в этом случае выбраны уровень пословной ошибки (word error rate, WER%) и процент случаев выбора правильной гипотезы (sentence error rate, SER%).

Результаты экспериментов отражены в табл. 1 и 2. В табл. 1 приведены перплексии всех используемых моделей. В табл. 2 даны результаты эксперимента по ранжированию – уровень пословной ошибки (WER%) и процент точности выбора правильной гипотезы (SER%).

Стоит отметить, что перплексии моделей, натренированных на лемматизиванном и нелемматизованном корпусе, строго говоря, не сравнимы по перплексии, поскольку количество неизвестных токенов, а значит и словарный состав корпусов различны: так, в корпусе словоформ оказалось много токенов «UNK», чем объясняется низкая перплексия этих моделей. Таким образом, важным обнадеживающим выводом, который можно сделать по данным табл. 2, является то, что модели на рекуррентных нейронных сетях демонстрируют существенно лучшие показатели в эксперименте, чем 5-граммная модель со сглаживанием Кнессера-Нея.

Рассмотрим теперь результаты эксперимента по ранжированию (табл. 3). Стоит сделать следующие замечания. Первое из них состоит в заметном превосходстве рекуррентных нейронных сетей над сглаженными *n*-граммами. Второй заметный факт – это противоречивое

Таблица 2. Перплексии моделей на тестовой выборке

Модель	Перплексия
$KN3_{lem}$	272,8
$KN4_{lem}$	272,2
$KN5_{lem}$	273
$KN3_{tok}$	128,72
$KN4_{tok}$	130,76
$KN5_{tok}$	132
RNN100	240,13
RNN200	230,45
RNN300	231
RNN400	231,87
RNN500	231,21

Таблица 3. Результаты моделей в эксперименте по ранжированию

Model	WER%	SER%
$KN5_{lem}$	16,62	40,8
$KN5_{tok}$	18,09	42,72
$KN5_{lem}$ + morph	15,58	43,98
KN_{lem} all	17,05	40,82
KN_{lem} all + morph	15,74	43,67
$KN_{lem+tok}$ all	15,74	39,24
$KN_{lem+tok}$ all + morph	15,89	43,35
all models	14,78	40,5
RNN100	17,55	43,67
RNN200	15,35	40,5
RNN300	17,09	43,98
RNN400	16,58	41,77
RNN500	17,43	43,67
RNN all	15,35	38,29
RNN all + morph	14,58	41,45

влияние морфологической модели на конечный результат: улучшение пословной ошибки при явной тенденции к голосованию за неверную гипотезу предложения. Это можно объяснить тем фактом, что оценка, возвращаемая морфологическим анализатором, пропорциональна вероятности лучшего разбора $P(tag_1^T | word_1^T)$. По этой причине данная оценка имеет тенденцию к выбору гипотез с наименьшей энтропией разбора. Стоит признать, что данная оценка не вполне подходит к решаемой нами задаче. Третий заметный факт состоит в несколько хаотичном характере результатов рекуррентных моделей: некоторые из них демонстрируют достаточно скромные результаты, однако их интерполяции обеспечивают наилучшие результаты.

Эксперименты по ранжированию в целом демонстрируют превосходство рекуррентных моделей. Наилучшая комбинация задействует оценку, возвращаемую морфологическим анализатором, и оценки, полученные от рекуррентных моделей. Таким образом, обеспечивается комбинирование морфологической и словарной информации. Данный результат свидетельствует о том, что риски в данном направлении могут быть продолжены.

Выводы

В статье был предложен простой эксперимент для проверки применимости рекуррентных нейронных сетей с внешним классификатором грамматических форм к русскому языку. В ходе эксперимента комбинировались отклики различных языковых моделей с целью ранжирования списка гипотез, возвращенных системой распознавания речи. Результаты указывают на то, что языковые модели на рекуррентных нейронных сетях превосходят результаты сглаженных *n*-граммных моделей как по перплексии, так и по уровню пословной ошибки. Тем не менее использование морфологии пока проблемно для рекуррентной нейронной сети. Следующим этапом работы может стать попытка тренировки классификатора с выпуклой функцией ошибки, где вычисленные векторы скрытого слоя нейронной сети будут служить в качестве признаков.

Список литературы

[1] Oparin I. *Language Models for Automatic Speech Recognition of Inflectional Languages*. *PhD thesis*. University of West Bohemia, Pilsen. 2008. P. 125.

[2] E.W.D. Whittaker. *Statistical Language Modeling for Automatic Speech Recognition of Russian and English*. *PhD Thesis*. Cambridge University. 2000. P. 141.

[3] Deoras A., Mikolov T., Kombrik S. Approximate inference: A sampling based modeling technique to capture complex dependencies in a language model, *Speech Communication, Speech Communication,* 2013.Vol. 55. № 1. P. 162.

[4] Bellegarda J. Exploiting latent semantic information in statistical language modeling, *Proceedings of the IEEE,* August, 2000. № 88. P. 1279.

[5] Gildea D., Hofmann T. Topic-based language models using EM. History. *Proceedings of the 6th European Conference on Speech Communication and Technology*, 1999. № 6. P. 2167.

[6] Andrieu C. et al. *An introduction to MCMC for machine learning. Machine learning,* 2003. Vol. 50. N 1-2. P. 5.

[7] Mikolov T. et al. Recurrent neural network based language model. *INTERSPEECH*. 2010. Vol. 2. P. 3.

[8] Mikolov T. *Statistical Language Models based on Neural Networks. PhD thesis*. Brno: University of Technology. 2012. P. 133.

[9] Vazhenina D., Markov K., Zelezny M. et al. Evaluation of Advanced Language Modeling Techniques for Russian LVCSR, *SPECOM 2013, LNAI 8113*. 2013. P. 124.

[10] Mikolov T. et al. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781,* 2013.

[11] Mikolov T., Sutskever I., Chen K., Corrado G. and Dean J. Distributed Representations of Words and Phrases and their Compositionality, *Proceedings of NIPS, URL: http://arxiv.org/pdf/1301.3781v3.pdf,* 2013.

[12] Elman J. Finding Structure in Time, *Cognitive Science,* 1990. № 14. P. 179.

[13] Bengio Y., Simard P., Frasconi P. Learning long-term dependencies with gradient descent is difficult, *Neural Networks, IEEE Transactions on,* 1994. Vol. 5. № 2. P. 157.

[14] Pascanu R., Mikolov T., Bengio Y. On the difficulty of training recurrent neural networks, *arXiv preprint arXiv:1211.5063,* 2012.

[15] Hochreiter S., Schmidhuber J. Bridging long time lags by weight guessing and "Long Short-Term Memory", *Spatiotemporal models in biological and artificial systems,* 1996. Vol. 37. P. 65.

[16] Muzychka S., Romanenko A., Piontkovskaja I. Conditional Random Field for morphological disambiguation in Russian, *Conference Dialog,* 2014. P. 11.

[17] Mikolov T., Kombrik S. RNNLM – Recurrent Neural Network Modeling Toolkit. *ASRU*, 2011. P. 456.

[18] Joachims T. Optimizing search engines using click through data, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining,* ACM. 2002. P. 133.