ИСПОЛЬЗОВАНИЕ РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РАНЖИРОВАНИЯ СПИСКА ГИПОТЕЗ В СИСТЕМАХ РАСПОЗНАВАНИЯ РЕЧИ

USE RECURRENT NEURAL NETWORKS FOR RANKING LIST HYPOTHESES IN SPEECH RECOGNITION SYSTEM

M. Kudinov

Annotation

The article presents the preliminary results of the use of recurrent neural networks for language modeling on Russian material. It solves the problem of ranking equally recognition hypotheses. To reduce the sparsity of data models were assessed for lemmatizovannom news package. It is also used to predict the morphological information. For the final sorting was used support vector for ranking. The article shows that the combination of neural networks and morphological model gives better results than a 5–gram model with smoothing Knessera–Ney.

Keywords: language model, a recurrent neural network, inflected languages, ranking the hypotheses, speech recognition.

Кудинов Михаил Сергеевич
Аспирант, Федеральный исследовательский центр ИУ РАН

Аннотация

В статье представлены предварительные результаты использования рекуррентных нейронных сетей для языкового моделирования на материале русского языка. Решалась задача ранжирования равновероятных гипотез распознавания. Для уменьшения разреженности данных модели оценивались на лемматизованном новостном корпусе. Также для предсказаний использовалась морфологическая информация. Для финальной сортировки была использован метод опорных векторов для ранжирования. В статье показано, что комбинация нейронных сетей и морфологической модели дает лучшие результаты, чем 5–граммная модель со сглаживанием Кнессера–Нея.

Ключевые слова:

Языковые модели, рекуррентная нейронная сеть, флективные языки, ранжирование гипотез, распознавание речи.

1. Введение

Известно, что проблема статистического моделирования флективных языков представляет большую сложность, чем для 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 году 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, использующей 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 работе было решено поставить предварительные эксперименты и решить более простую задачу, а именно произвести переранжирование гипотез распознавания, исходя из оценок отдельной лексической модели, основанной на рекуррентной нейронной сети, и морфологической модели, основанной на условных случайных полях.

Статья организована следующим образом. В разделе 2 приводится общая информация о рекуррентных нейронных сетях. В разделе 3 обсуждается применимость оригинальной архитектуры рекуррентной нейронной сети к статистическому моделированию флективных языков и сопутствующим проблемам. В разделах 4 и 5 описаны результаты экспериментов.

2. Рекуррентная нейронная сеть для статистического моделирования языка

Рекуррентные нейронные сети впервые были рассмотрены в [12] Элманом в 1990 году. 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 года, 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 ?M. Каждое входное слово (допустим, с индексом l) в словаре объемом $|L|$ слов представляется в виде $|L|$–мерного вектора $w = < 01,...,1_l,0_{l+1},...,0_{|L|} >$ 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 использована в [7] Т.Миколовым, который использовал рекуррентную сеть Элмана для предсказания слов по контексту. Результирующая модель описывалась следующими уравнениями:

$$P\left(w_k \left| w_{t-1}, h_{t-1}\right.\right) = y_{wk}\left(t\right) \quad (1)$$

$$y\left(t\right) = s\left(V \cdot h_t\right) \quad (2)$$

$$h_t = \sigma\left(U \cdot x + W \cdot h_{t-1}\right) \quad (3)$$

где логистическая функция активации (4), а софтмакс–функция (5).

$$\sigma\left(x\right) = \frac{1}{1+e^{-x}} \quad (4)$$

$$s\left(x\right) = \frac{e^x}{\sum_i e^{x_i}} \quad (5)$$

x_t – вектор с единственной единичной координатой; ht – рекуррентный слой; y – выходной слой, где каждому k–му элементу соответствует вероятность $\mathrm{P}(wk|w_{t-1},h_{t-1})$, $W_{H \times H}$ 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.

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 , $k < t$, отражающего влияние предыдущих значений на скрытом слое на последующие, стремится к нулю (или к бесконечности) с экспоненциальной скоростью по (t - k) [13], [14]:

$$\frac{\partial h_t}{\partial h_k} = \prod_{k<i\le t} \frac{\partial h_i}{\partial h_{i-1}} = \prod_{k<i\le t} W^T diag\left(\sigma'\left(h_{i-1}\right)\right), \quad (6)$$

где

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

В работах [13], [14], [15] было показано, что в зависимости от свойств матрицы W значение выражения (6) 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.

3. Применимость подхода к моделированию флективных языков

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.

Чтобы обойти эту проблему, можно было бы использовать схему на **рисунке 1**. 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 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.

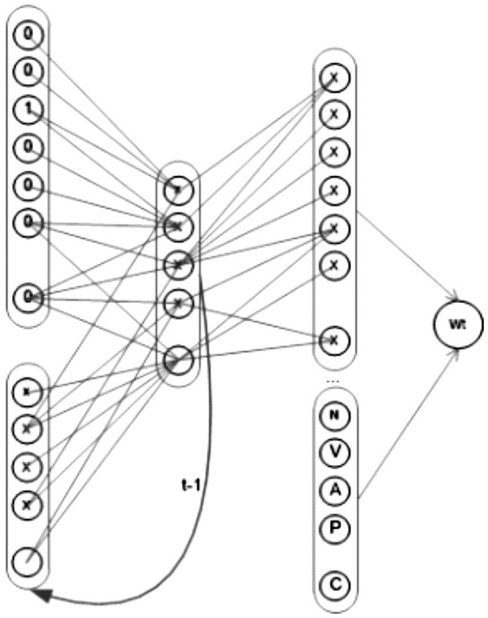

Рис. 1. Рекуррентная нейронная сеть с внешним классификатором.

В данной статье рассматривается предварительный эксперимент, целью которого является проверка гипотезы о том, что комбинация нейронных сетей, обученных на леммах, дает лучший результат, чем комбинация n–граммных моделей с дисконтированием Кнессера–Нея. Его успешность может косвенно указать на перспективы архитектуры на рисунке 1.

4. Описание экспериментов

Для оценивания моделей было поставлено два эксперимента. В первом из них оценивалась перплексия моделей на новостном корпусе. Во втором эксперименте оценивалось влияние выбора языковой модели на уровень пословной ошибки системы распознавания речи.

Перплексией выборки D в модели θ называется величина, определяемая как:

$$P(D;\theta) = 2^{\frac{1}{|D|}\sum_{w \in D} \log_2 p(w|h;\theta)},$$

где h – левый контекст слова w.

Пословная ошибка системы распознавания речи вычисляется как:

$$WER = \frac{I + D + R}{N},$$

где I, D, R – соответственно количества вставок, удалений и замен слов в распознанном тексте, относительно исходного, а N – длина исходного текста в словах.

В качестве базовой модели для оценки была выбрана n–граммная модель со сглаживанием Кнессера–Нея [17] (далее KNn, где n – порядок модели):

$$P_{KN}(w_i|h) = \begin{cases} \dfrac{C(hw_i) - D}{C(h)} \\ \alpha(w_{i-1}) \dfrac{\left|\left\{\bar{h} : C(\bar{h}w_i) > 0\right\}\right|}{\sum_w \left|\left\{\bar{h} : C(\bar{h}w) > 0\right\}\right|} \end{cases},$$

где D – настраиваемый на валидационной выборке параметр, α – нормирующий коэффициент, h – (n-1)–граммный левый контекст слова wi.

N –граммная модель со сглаживанием Кнессера–Нея является наиболее эффективной сглаженной n–граммной моделью [17]. По этой причине она была выбрана в качестве базовой.

В ходе эксперимента были рассмотрены KNn модели, натренированные как на леммах, так и на словоформах. По этой причине, было подготовлено 2 обучающих корпуса.

Рекуррентные модели и KNn на леммах были натренированы на новостном корпусе объемом приблизительно в $2 \cdot 10^6$ словоупотреблений (токенов). Примерно по 10% предложений было выделено для валидационной и тестовой выборок. Каждый текст был обработан морфологическим анализатором / лемматизатором для русского языка [16] со встроенным словарем примерно в $2 \cdot 10^6$ словоформ.

Все леммы, не входящие в 10000 наиболее частотных, заменялись на токен "UNK". Таким образом, словарь модели составил 10000 лемм.

KNn на словоформах были натренированы на корпусе, который был получен заменами на "UNK" тех токенов, которые не принадлежали леммам словаря корпуса.

На тестовых частях двух полученных корпусов проводились также эксперименты по определению перплексии.

Во втором эксперименте оценивалось влияние моделей на ранжирование гипотез распознавания. Гипотезы генерировались внешней системой распознавания фирмы Nuance. К сожалению, декларируемый показатель WER данной системы, как и полный список гипотез, недоступен. Использовался русскоязычный корпус предложений со студийным качеством записи и транскрипциями. Аудиофайлы подавались на вход системе распознавания. На выходе получалось до 10 гипотез. В результате была получена 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 рекуррентных 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 сетей 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 использовалась модель ranking SVM [18], где в качестве признаков использовались оценки моделей. Результирующая модель обучалась ранжированию гипотез в списке на 2 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%).

5. Результаты

Результаты экспериментов приводятся в таблицах 1 и 2.

В **таблице 1** приведены перплексии всех используемых моделей.

В **таблице 2** приведены результаты эксперимента по ранжированию – уровень пословной ошибки (WER%) и процент точность выбора правильной гипотезы (SER%).

Стоит отметить, что перплексии моделей, натренированных на лемматизиванном и нелемматизованном кор-

Таблица 1.

Перплексии моделей на тестовой выборке.

Модель	Перплексия	Модель	Перплексия
$KN3_{lem}$	272.8	RNN100	240.13
$KN4_{lem}$	272.2	RNN200	230.45
$KN5_{lem}$	273	RNN300	231
$KN3_{tok}$	128.72	RNN400	231.87
$KN4_{tok}$	130.76	RNN500	231.21
$KN5_{tok}$	132		

Таблица 2.

Результаты моделей в эксперименте по ранжированию.

Model	WER %	SER %	Model	WER %	SER %
$KN5_{lem}$	16.62	40.8	RNN100	17.55	43.67
$KN5_{tok}$	18.09	42.72	RNN200	15.35	40.5
$KN5_{lem}$ + morph	15.58	43.98	RNN300	17.09	43.98
KN_{lem} all	17.05	40.82	RNN400	16.58	41.77
KN_{lem} all + morph	15.74	43.67	RNN500	17.43	43.67
$KN_{lem+tok}$ all	15.74	39.24	RNN all	15.35	38.29
$KN_{lem+tok}$ all + morph	15.89	43.35	RNN all + morph	14.58	41.45
all models	14.78	40.5			

пусе, строго говоря, не сравнимы по перплексии, поскольку количество неизвестных токенов, а значит и словарный состав корпусов, различны. Этим фактом объясняется и низкая перплексия моделей на словоформах: количество токенов "UNK" в корпусе было велико. Таким образом, важным обнадеживающим выводом, который можно сделать из приведенной таблицы, 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. Стоит сделать следующие замечания.

Первое из них состоит в заметном превосходстве рекуррентных нейронных сетей над сглаженными n-граммами.

Второй заметный факт – это противоречивое 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 $P(tag_1^T \mid word_1^T)$. 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 результаты в данном направлении могут быть продолжены.

Причину эффективности рекуррентной архитектуры для моделирования языка можно объяснить предположением, высказанным в [8]. Его суть состоит в том, что более высокие по сравнению с KNn результаты, демонстрируемые рекуррентной моделью, обусловлены ее способностью к кластеризации близких по смыслу контекстов. KNn фактически предоставляет лишь возможность "возврата" (backing-off) к модели меньшего порядка в случае ненадежной оценки вероятности n-граммы по корпусу.

Таким образом, KNn игнорирует синонимию контекстов и, в конечном счете, требует большего количества обучающих данных.

6. Заключение

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, эксперименты должны быть продолжены в двух направлениях: проверка воспроизводимости результатов при наличии большей обучающей выборки; и конструирование языковой модели на рекуррентной нейронной сети для предсказания словоформ русского языка.

ЛИТЕРАТУРА

1. Oparin: Language Models for Automatic Speech Recognition of Inflectional Languages. PhD thesis, University of West Bohemia, Pilsen, 2008.
2. E.W.D. Whittaker: Statistical Language Modeling for Au– tomatic Speech Recognition of Russian and English. PhD Thesis, Cambridge University, 2000.
3. A.Deoras, T.Mikolov, S. Kombrik: Approximate inference: A sampling based modeling technique to capture complex dependencies in a language model. Speech Communication, 2012
4. J.Bellegarda: Exploiting latent semantic information in statistical language modeling. Proc. IEEE. 88, 2000
5. D.Gildea, T.Hoffman: Topic–Based Language Models Using EM. In Proceedings of EUROSPEECH, 1999
6. Y.Bengio, R.Ducharme, P.Vincent, C.Jauvin: A Neural Probabilistic Language Model. Journal of machine learning research, 2003
7. T.Mikolov, M.Karafiat, L.Burget, J.Cernocky, S.Khudanpur: Recurrent neural network based language model, In: Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010), Makuhari, Chiba, JP
8. T.Mikolov,: Statistical Language Models based on Neural Networks. PhD thesis, Brno University of Technology, 2012.
9. D.Vazhenina, K.Markov, Evaluation of Advanced Language Modeling Techniques for Russian LVCSR, M.Zelezny et al. (Eds.): SPECOM2013, LNAI 8113, pp.124–131, 2013.
10. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.
11. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.
12. J.Elman. Finding Structure in Time. Cognitive Science, 14, 179–211, 1990.
13. Y.Bengio, P.Simard, P.Frasconi. Learning Long–Term Dependencies with Gradient Descent is Difficult, IEEE Transactions on neural networks, 1994
14. R.Pascanu, T.Mikolov, Y.Bengio. On the difficulty of training Recurrent Neural Networks, CoRR, 2012
15. Hochreiter, S. and Schmidhuber, J. (1996). Bridging long time lags by weight guessing and Long Short–Term Memory. In F.Silva, J.Principe, L.Almeida, Spatiotemporal models in biological and artificial systems
16. S.Muzychka, A.Romanenko, I.Piontkovskaja. Conditional Random Field for morphological disambiguation in Russian., Conference Dialog–2014, Bekasovo, 2014
17. J. Goodman. A Bit of Progress in Language Modeling, Microsoft Research Technical Report, 2001
18. T. Joachims. Optimizing Search Engines using Clickthrough Data, Proceedings of the ACM Conference on Knowledge Discovery and Data Mining, 2003