Языковое моделирование для систем автоматического распознавания слитной русской речи

Language Modelling for Continuous Russian Speech Recognition Systems

И.С. Кипяткова [1]
Irina Kipyatkova [1]
kipyatkova@iias.spb.su

А.А. Карпов [2]
Alexey Karpov [2]
karpov@iias.spb.su

[1] СПИИРАН, ГУАП
Санкт-Петербург, Российская Федерация

[1] SPIIRAS, SUAI
Saint Petersburg, Russian Federation

[2] СПИИРАН, Университет ИТМО
Санкт-Петербург, Российская Федерация

[2] SPIIRAS,
ITMO University
Saint Petersburg, Russian Federation

Abstract

The paper deals with a research of three types of language models, namely trigram, factored, and neural network language models. Factored language models were created using several linguistic factors (word, lemma, stem, part-of-speech, and morphological tag). Recurrent neural network based language models were created with different number of hidden units. Experiments on application of created models in the continuous Russian speech recognition system were conducted. Trigram language model was used at decoding stage, factored and neural network models were used for N-best list rescoring. The relative word error rate reduction of 8% using factored language models and 14% relative word error rate reduction using neural network language model with respect to the baseline model was achieved.

Keywords: *Language modeling, factored language model, neural network language model, Russian speech recognition systems*

Аннотация

Аннотация: Статья посвящена исследованию трех типов статистических моделей языка: триграммной, факторной и нейросетевой. Были созданы факторные модели русского языка с использованием различных лингвистических факторов (словоформа, лемма, основа слова, часть речи и метка морфологических признаков) и модели на базе рекуррентных искусственных нейронных сетей с различным числом элементов в скрытом слое. Проведены эксперименты по применению созданных моделей в системе распознавания слитной русской речи с большим словарем, при этом триграммная модель использовалась на этапе декодирования, а факторная и нейросетевая – на этапе переоценки списков лучших гипотез распознавания. Относительное уменьшение количества неправильно распознанных слов составило 8% при использовании факторной модели языка и 14% – при использовании нейросетевой модели языка.

Ключевые слова: *Языковое моделирование, факторная модель языка, нейросетевая модель языка, системы распознавания русской речи*

1 Введение

Модель языка является необходимым компонентом систем автоматического распознавания слитной речи. Наиболее распространенной моделью языка является статистическая модель на основе n-грамм слов, позволяющая оценить вероятность появления цепочки из n слов в некотором тексте. Данная модель показала свою эффективность для языков с жестким порядком слов (например, английского), однако русский язык обладает рядом особенностей, снижающих эффективность статистических моделей. В русском языке практически свободный порядок слов, кроме того, русский язык является синтетическим флективным языком с богатой морфологией, что приводит к существенному увеличению размера словаря системы распознавания, а также к увеличению коэффициента неопределенности (perplexity) n-граммных моделей языка. В данной статье приводится описание моделей языка, позволяющих учитывать особенности русского языка.

2 Разновидности моделей языка

Существует несколько разновидностей статистических моделей языка, позволяющих моделировать длинный контекст или дальнодействующие связи между словами. Одной из таких разновидностей являются триггерные модели (Trigger models). В этом методе появление инициирующего слова в истории увеличивает вероятность другого слова, называемого целевым, с которым оно связано [Vaičiūnas 2006].

Упрощенной версией триггерных пар является кэш-модель (cache model), которая увеличивает вероятность появления слова в соответствии с тем, как часто данное слово встречалось в истории, поскольку считается, что, употребив конкретное слово, диктор будет использовать его еще раз, поскольку либо оно является характерным для конкретной темы, либо потому что диктор имеет тенденцию использовать это слово в своем лексиконе [Vaičiūnas 2006].

В работе [Protasov 2008] предлагается дальнодействующая триграммная модель, представляющая собой триграммную модель, которая предсказывает вероятность

появления слова не только по непосредственно предшествующим словам, но и по словам, находящемся на большем расстоянии от предсказываемого слова. Лежащая в основе "грамматика" представляет собой множество пар слов, которые могут быть связаны вместе через несколько разделяющих слов.

Еще одним типом модели языка, позволяющим моделировать дальнодействующие связи в предложении, является синтаксическо-статистическая модель, предложенная в работе [Karpov 2014]. Для создания такой модели вначале выполняется статистический анализ текстового корпуса и создается список n-грамм слов. Затем производится синтаксический анализ, в ходе которого выявляются грамматически связанные пары слов (синтаксические группы), которые были разделены в тексте другими словами. Такие синтаксические группы добавляются к списку n-грамм слов, полученных в ходе статистического анализа текстового корпуса. Диаграмма создания синтаксическо-статистической модели языка показана на рисунке 1.

Рис. 1: Процесс создания синтаксическо-статистической модели языка

Модели, основанные на частях слов, (Particle-based models) используются для языков с богатой морфологией, например, флективных языков [Vaičiūnas 2006]. В этом случае слово w разделяется на некоторое число ($L(w)$) частей (морфем) с помощью функции $U : w \to u^1, u^2, \ldots, u^{L(w)}, u^i \in \Psi$, где Ψ – это набор частиц слова. Существует два типа методов разделения слов на морфемы: словарные и алгоритмические [Kurimo 2009]. Преимуществом алгоритмических методов является то, что они опираются лишь на анализ текста и не используют никаких дополнительных знаний, что позволяет анализировать текст на любом языке, но при этом слова разбиваются на псевдоморфемные единицы. Преимуществом словарных методов является то, что они позволяют получить правильное разбиение слов на морфемы, что может быть использовано далее на уровне пост-обработки гипотез распознавания фраз.

Еще одной моделью, которая может быть использована для языков с богатой морфологией, является факторная модель языка (ФМЯ), которая впервые была предложена в работе [Bilmes 2003] для моделирования арабского языка. Эта модель объединяет различные признаки слова (факторы), при этом слово представляется как вектор k факторов $Y_i = (F_i^1, F_i^2, \ldots, F_i^k)$. В качестве факторов могут использоваться: словоформа, часть речи, основа, корень слова и другие морфологические и грамматические признаки.

Модель языка может быть построена на базе искусственных нейронных сетей, в частности для данной задачи успешно используются рекуррентные ИНС (РИНС), которые в первые были предложены в работе [Elman 1990]. Применение РИСН для

моделирования языка представлено в работе [Mikolov 2010]. Преимущество данной модели состоит в том, что скрытый слой хранит весь контекст, предшествующий рассматриваемому слову. Сеть имеет входной слой, скрытый слой (также называемый контекстным слоем или состоянием) и выходной слой. Выходной слой после обучения нейронной сети представляет собой вероятностное распределение следующего слова при данном предыдущем слове и состоянии скрытого слоя в предшествующий временной шаг. Размер скрытого слоя обычно выбирается эмпирически.

В данной статье приведено описание процесса создания факторной и нейросетевой модели для русского языка и применения таких моделей на этапе переоценки лучших гипотез распознавания (N-best list) в системах автоматического распознавания слитной речи.

3 Триграммная модель русского языка

В ходе нашего исследования для создания моделей языка был собран и автоматически обработан текстовый корпус, сформированный из интернет-сайтов ряда электронных газет. Общий объем собранного корпуса после его обработки составил свыше 350 млн словоупотреблений, корпус содержит около 1 млн уникальных словоформ [Kipyatkova 2010]. Данный корпус был использован для обучения как базовой (триграммной) модели языка, так и факторных и нейросетевых моделей. Для создания триграммной модели языка использовался пакет программных средств SRILM [Stolcke 2011]. В ходе предыдущих экспериментов по автоматическому распознаванию слитной русской речи с применением различных моделей языка было определено, что наименьшее количество неправильно распознанных слов достигается при применении модели языка со словарем 150 тыс. словоформ [Kipyatkova 2013]. Оценка созданной модели языка (определение коэффициента неопределенности) осуществлялась по текстовому корпусу, в который вошли материалы интернет-газеты, не используемой в обучающем корпусе. Коэффициент неопределенности триграммной модели языка был равен 553.

4 Факторная модель русского языка

Морфологический анализ обучающего текстового корпуса производился с помощью программы “VisualSynan” проекта AOT [Sokirko 2004]. В работе были использованы 5 лингвистических факторов: словоформа, лемма, основа, часть речи, метка морфологических признаков. Таким образом, все слова в обучающем текстовом корпусе были заменены на факторы. Например, словоформа “схеме” была заменена на следующую последовательность факторов: “W-схеме L-схема S-схем P-сущ M-bc”, где W – словоформа, L – лемма, S – основа, P – часть речи, M – метка морфологических признаков, содержащая всю грамматическую информацию о слове (в данном примере она означает, что словоформа является существительным женского рода в единственном числе и дательном падеже). По обучающему текстовому корпусу были созданы двухфакторные модели, включающие словоформу и один из других перечисленных выше факторов.

При создании статистических моделей языка для решения проблемы недостатка обучающих данных используется методика возврата (back-off) [Moore 2001], суть

которой состоит в том, что, когда некоторая n-грамма отсутствует в обучающем текстовом корпусе или частота ее появления очень низкая, то вместо нее используется вероятность (n-1)-граммы, умноженная на коэффициент возврата. В n-граммных моделях возврат осуществляется путем отбрасывания сначала наиболее дальнего слова, затем второго по дальности слова и т.д. Для факторных моделей языка с двумя факторами возможны два варианта пути возврата: 1) сначала опускаются дальняя словоформа и фактор, затем – ближняя словоформа и фактор (см. рис. 2а); 2) сначала опускаются словоформы в порядке их удаленности от рассматриваемой, а затем факторы в том же порядке (см. рис. 2б). Подробно процесс создания факторной модели русского языка описан в работе [Kipyatkova 2015a].

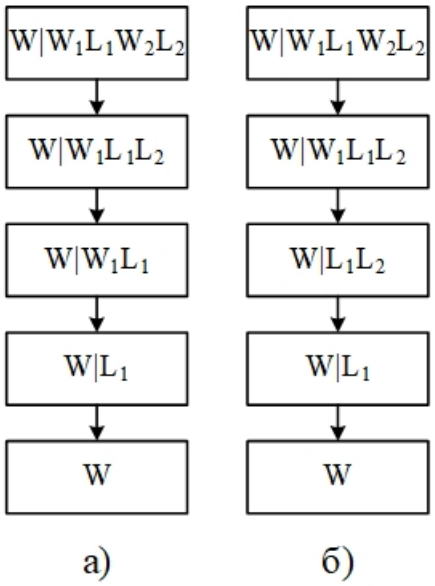

Рис. 2: Фиксированные пути возврата ФМЯ с факторами слово и лемма: а) первый путь возврата; б) второй путь возврата.

В таблице 1 показаны коэффициенты неопределенности для созданных ФМЯ. У моделей, имеющих путь возврата 1, значение коэффициента неопределенности оказалось меньше независимо от используемых факторов. Наименьшее значение коэффициента неопределенности имеет модель с факторами “словоформа” и “лемма”, имеющая путь возврата 1.

5 Нейросетевая модель русского языка

Обучение РИНС осуществлялось с помощью свободно доступного программного модуля RNNLM toolkit (Recurrent Neural Network Language Modeling Toolkit) [Mikolov 2010]. Для сокращения скорости обучения нейронной сети была выполнена факторизация выходного слоя. Слова были разбиты на классы в соответствии с их частотой. Вначале вычислялось распределение вероятности для классов, затем – распределение вероятности для слов, которые относятся к соответствующему классу. Были созданы модели с числом элементов в скрытом слое, равном 100, 300 и

Таблица 1: Коэффициенты неопределенности ФМЯ

Факторы	Коэффициент неопределенности	
	Путь 1	Путь 2
WM	566	691
WL	529	577
WP	623	729
WS	595	672

500 и числом классов, равном 100 и 500 [Kipyatkova 2015b]. Значения коэффициента неопределенности для созданных моделей представлены в таблице 2.

Таблица 2: Коэффициенты неопределенности для нейросетевых моделей русского языка

Количество классов	Количество элементов в скрытом слое		
	100	300	500
100	981	997	7661
500	1074	843	870

6 Исследование статистических моделей языка для распознавания слитной русской речи

Акустические модели, в качестве которых использовались лево-правые скрытые марковские модели с тремя состояниями, создавались с помощью инструментария HTK (Hidden Markov Model Toolkit) [Young 2009] по речевому корпусу слитной русской речи, содержащему записи 50 дикторов – носителей русского языка. Общий объем корпуса составляет 13,5 Гб, длительность записей – более 21 ч. Для тестирования системы распознавания речи использовался корпус, содержащий 100 слитно произнесенных фраз, взятых из материалов интернет-газеты “Фонтанка.ru”. Каждая фраза была произнесена 5 дикторами. Длина каждого предложения составляет от 6 до 20 слов. Сессия записи для каждого диктора длилась от 20 до 40 мин, при этом чистая речь составила 15-30 мин. Общий объем тестового корпуса – 200 Мб аудиоданных. Система автоматического распознавания слитной русской речи была построена на основе декодера Julius 4.2 [Lee 2009]. Оценка работы системы распознавания проводилась по критерию качества распознавания речи с использованием показателя процента неправильно распознанных слов в речи (WER — Word Error Rate). На этапе декодирования речи применялась триграммная модель языка, при этом было получено WER=26,54%, и были созданы списки лучших гипотез распознавания с числом гипотез, равным 10, 20, 50. Затем была выполнена переоценка гипотез с

использованием созданных факторных и нейросетевых моделей языка.

В таблице 3 представлены результаты распознавания, полученные после переоценки списков лучших гипотез распознавания факторной моделью языка, а также факторной моделью, интерполированной с базовой моделью языка с разными коэффициентами интерполяции. Под интерполяцией моделей языка понимается линейная комбинация вероятностей слов, полученных от разных моделей, с учетом весовых коэффициентов каждой модели. Коэффициент интерполяции, равный 1, означает, что использовалась только факторная модель языка.

Таблица 3: Количество неправильно распознанных слов (WER, %) после переоценки списков лучших гипотез распознавания ФМЯ

Модель языка	N=10		N=20		N=50	
	Путь 1	Путь 2	Путь 1	Путь 2	Путь 1	Путь 2
WM+3-гр.	24,83	24,94	**24,44**	24,78	24,55	24,66
WL+3-гр.	25,79	25,71	25,58	25,43	25,60	25,37
WP+3-гр.	25,43	25,54	25,07	25,24	25,15	25,26
WS+3-гр.	25,82	26,01	25,88	25,90	25,90	26,10

Наименьший процент неправильно распознанных слов (WER=24,44%) был получен при переоценке списка из 20 гипотез факторной моделью, в которой в качестве факторов использовались словоформа и морфологический таг, созданной с путем возврата 1, интерполированной с триграммной моделью.

Для экспериментов по применению нейросетевых моделей языка использовались те же списки лучших гипотез распознавания, что и для экспериментов с факторными моделями. Также была выполнена интерполяция нейросетевых и триграммных моделей языка. Полученные результаты представлены в таблице 4.

Из таблицы видно, что применение нейросетевой модели языка позволило сократить ошибку распознавания слов за исключением применения РИНС со 100 элементами в скрытом слое без интерполяции с триграммной моделью. Использование РИНС с числом классов, равным 100, дало лучшие результаты распознавания, чем применение РИНС с числом классов, равным 500. Наилучший результат (WER=22,87%) был получен при применении модели языка на основе РИНС с 500 элементами в скрытом слое и числом классов, равным 100, интерполированной с триграммной моделью с коэффициентом интерполяции, равным 0,5.

7 Заключение

Проведено исследование трех типов статистических моделей русского языка (триграммной, факторной и неросетевой) для систем автоматического распознавания слитной русской речи. Преимущество ФМЯ по сравнению со статистическими n-граммными моделями состоит в том, что в модель языка включается дополнительная лингвистическая информация, что позволяет повысить качество распознавания речи для языков с богатой морфологией, в том числе и русского. Преимуществом моделей на базе РИНС является то, что они могут хранить языковой контекст произволь-

Таблица 4: Количество неправильно распознанных слов (WER, %), полученное после переоценки различных списков N лучших гипотез

Модель языка	Коэффициент интерполяции, λ	N=10	N=20	N=50
РИНС с 100 элементами в скрытом слое + триграммная МЯ	1,0	26,33	26,65	26,72
	0,6	25,13	25,06	24,98
	0,5	25,13	24,89	24,91
	0,4	25,06	24,72	24,72
РИНС с 300 элементами в скрытом слое + триграммная МЯ	1,0	25,41	25,30	25,49
	0,6	24,68	24,53	24,51
	0,5	24,59	24,04	24,18
	0,4	24,53	23,97	24,10
РИНС с 500 элементами в скрытом слое + триграммная МЯ	1,0	24,51	23,67	23,97
	0,6	23,76	23,07	22,96
	0,5	23,65	23,00	**22,87**
	0,4	23,82	23,26	23,24

ной длины. Проведенные эксперименты по распознаванию слитной русской речи, в которых разработанные факторные и нейросетевые модели использовались на этапе переоценки лучших гипотез распознавания, показали, что применение таких моделей позволяет снизить процент неправильно распознанных слов. При этом относительное уменьшение количества неправильно распознанных слов составило 8% при использовании факторной модели и 14% – при использовании нейросетевой модели языка по сравнению с результатами, полученными при применении триграммной модели языка.

8 Благодарности

Работа выполнена при финансовой поддержке фонда РФФИ (проекты № 18-07-01216 и 18-07-01407) и Совета по грантам Президента РФ (проекты № МК-1000.2017.8 и МД-254.2017.8).

Список литературы

[Kipyatkova 2010] Kipyatkova, I.S., Karpov, A.A. (2010). Avtomaticheskaja obrabotka i statisticheskij analiz novostnogo tekstovogo korpusa dlja modeli jazyka sistemy raspoznavanija russkoj rechi [Automatic Processing and Statistic Analysis of News Text Corpus for a Language Recognition System Model of Russian Speech]. Informacionno-upravljajushhie sistemy [Information-control Systems], № 4(47). Pp. 2–8 (In Russian) = Кипяткова И.С., Карпов А.А. Автоматическая обработка и статистический анализ новостного текстового корпуса для модели языка систе-

мы распознавания русской речи // Информационно-управляющие системы, № 4(47), 2010. С. 2–8.

[Protasov 2008] Protasov, S.V. (2008). Vyvod i ocenka parametrov dal'nodejstvujushhej trigrammnoj modeli jazyka [Inference and Estimation of a Long-Range Trigram Model]. Materialy mezhdunarodnoj konferencii "Dialog 2008"[Proceedings of International Conference "Dialog 2008"]. S. 444-448 (in Russian) = Протасов С.В. Вывод и оценка параметров дальнодействующей триграммной модели языка. Материалы международной конференции "Диалог 2008". Москва, 2008. С. 444–448

[Sokirko 2004] Sokirko, A.V. (2004). Morfologicheskie moduli na sajte www.aot.ru [Morphological modules on the website www.aot.ru]. Trudy Mezhdunarodnoj konferencii "Dialog-2004"[Proceedings of International Conference "Dialog 2004"]., pp. 559-564 (in Russin) = Сокирко А.В. Морфологические модули на сайте www.aot.ru // Труды Международной конференции "Диалог-2004", 2004. С. 559–564.

[Bilmes 2003] Bilmes J. A., Kirchhoff K.(2003) Factored language models and generalized parallel backoff // Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Stroudsburg, PA, USA, 2003, Vol. 2. Pp. 4—6.

[Elman 1990] Elman, J.L. Finding (1990) Structure in Time // Cognitive Science, 1990, Vol. 14. Pp. 179—211.

[Karpov 2014] Karpov A., Markov K., Kipyatkova, I., Vazhenina, D., Ronzhin A. (2014) Large vocabulary Russian speech recognition using syntactico-statistical language modeling // Speech Communication, 2014, Vol. 56. Pp. 213-228 :http://dx.doi.org/10.1016/j.specom.2013.07.004.

[Kipyatkova 2013] Kipyatkova I., Karpov, A. (2013) Lexicon Size and Language Model Order Optimization for Russian LVCSR // Springer International Publishing Switzerland. M. Zelezny et al. (Eds.): SPECOM 2013, LNAI 8113. 2013. Pp. 219–226.

[Kipyatkova 2015a] Kipyatkova I. Karpov A. (2015a) Development of Factored Language Models for Automatic Russian Speech Recognition // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 27 — 30 мая 2015 г.)[Computational linguistics and intelligent technologies: Proceedings of International Conference "Dialog" (Moscow, 27 -30 may 2015]., Вып. 14, Москва, 2015a, С. 234–246.

[Kipyatkova 2015b] Kipyatkova I. Karpov A. (2015b) Recurrent Neural Network-based Language Modeling for an Automatic Russian Speech Recognition System // Proceedings of International Conference AINL-ISMW FRUCT, 2015b. Pp. 33–38.

[Kurimo 2009] Kurimo M., Hirsimäki T., Turunen V.T., Virpioja S., Raatikainen N. (2009) Unsupervised decomposition of words for speech recognition and retrieval // Proceedings of 13-th International Conference "Speech and Computer"SPECOM'2009, St. Petersburg, 2009. Pp. 23–28.

[Lee 2009] Lee A., Kawahara T. (2009) Recent Development of Open-Source Speech Recognition Engine Julius // Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2009). 2009. Pp. 131—137.

[Mikolov 2010] Mikolov T., Karafiát M., Burget L., Černocký J., Khudanpur S. (2010) Recurrent neural network based language model // Proceedings of INTERSPEECH'2010. 2010. Pp. 1045—1048.

[Mikolov 2011] Mikolov T., Kombrink S., Deoras A., Burget L., Černocký J. (2011) RNNLM - Recurrent Neural Network Language Modeling Toolkit // Proceedings of the 2011 ASRU Workshop, 2011. Pp. 196–201.

[Moore 2001] Moore G.L. (2001) Adaptive Statistical Class-based Language Modelling, PhD thesis, Cambridge University, 2001, – 193 p.

[Stolcke 2011] Stolcke A., Zheng J., Wang W., Abrash V. (2011) SRILM at Sixteen: Update and Outlook // Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop ASRU'2011.

[Vaičiūnas 2006] Vaičiūnas A. (2006) Statistical Language Models of Lithuanian and Their Application to Very Large Vocabulary Speech Recognition. PhD thesis. Vytautas Magnus University. Kaunas. –35 c.

[Young 2009] Young, S. et al. (2009). The HTK Book (for HTK Version 3.4). Cambridge, UK. –375 p.